Spatio-temporal modelling of climate-sensitive disease risk: towards an early warning system for dengue in Brazil

Submitted by

Rachel Lowe

to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Mathematics, September 2010.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

. . . . . .

Rachel Lowe

Abstract

The transmission of many infectious diseases is affected by climate variations, particularly for diseases spread by arthropod vectors such as malaria and dengue. Previous epidemiological studies have demonstrated statistically significant associations between infectious disease incidence and climate variations. Such research has highlighted the potential for developing climate-based epidemic early warning systems. To establish how much variation in disease risk can be attributed to climatic conditions, non-climatic confounding factors should also be considered in the model parameterisation to avoid reporting misleading climate-disease associations. This issue is sometimes overlooked in climate related disease studies. Due to the lack of spatial resolution and/or the capability to predict future disease risk (e.g. several months ahead), some previous models are of limited value for public health decision making.

This thesis proposes a framework to model spatio-temporal variation in disease risk using both climate and non-climate information. The framework is developed in the context of dengue fever in Brazil. Dengue is currently one of the most important emerging tropical diseases and dengue epidemics impact heavily on Brazilian public health services. A negative binomial generalised linear mixed model (GLMM) is adopted which makes allowances for unobserved confounding factors by including spatially structured and unstructured random effects. The model successfully accounts for the large amount of overdispersion found in disease counts. The parameters in this spatio-temporal Bayesian hierarchical model are estimated using Markov Chain Monte Carlo (MCMC). This allows posterior predictive distributions for disease risk to be derived for each spatial location and time period (month/season). Given decision and epidemic thresholds, probabilistic forecasts can be issued, which are useful for developing epidemic early warning systems.

The potential to provide useful early warnings of future increased and geographically specific dengue risk is investigated. The predictive validity of the model is evaluated by fitting the GLMM to data from 2001-2007 and comparing probabilistic predictions to the most recent out-of-sample data in 2008-2009. For a probability decision threshold of 30% and the pre-defined epidemic threshold of 300 cases per 100,000 inhabitants, successful epidemic alerts would have been issued for 94% of the 54 microregions that experienced high dengue incidence rates in South East Brazil, during February - April 2008.

Acknowledgements

This work was supported by the EUROBRISA network project (F/00 144/AT) kindly funded by the Leverhulme Trust. I have been extremely fortunate to be supervised by two excellent professors: David Stephenson and Trevor Bailey. I am grateful to David for his sound advice, encouragement and enthusiasm in my work throughout my PhD training. I would like to thank Trevor for always making available his support to help solve pressing dilemmas over the years. I would also like to express my gratitude to my Met Office co-supervisor Richard Graham for his thoughtful input. Thanks to Richard and Richard Betts, the Met Office kindly funded me to attend the 2009 Summer Institute on Climate Information for Public Health at the International Research Institute for Climate and Society, New York. This course greatly enhanced my knowledge of the subject area.

This project has greatly benefited from invaluable knowledge and data contributions from my Brazilian colleagues: Caio Coelho, Marilia Sá Carvalho, Christovam Barcellos, Evangelina Xavier Gouveia de Oliveira and Miguel Antonio Vieira Monteiro. I would also like to thank Madeleine Thomson, Tony Barnston, Sari Kovats, Helen Gurgel and Aidan Slingsby for helpful discussions during my PhD career. I am grateful to the organisers of the StatGIS09 conference for inviting me to submit an extended version of my presentation for a special issue on 'GeoInformatics for environmental surveillance' for publication in Computers & Geosciences (Lowe et al., 2010).

I have received excellent support and guidance from my colleagues in Exeter Climate Systems. Tim Jupp greatly contributed to the visualisation component of this research. I am extremely grateful for his exciting ideas and provision of R functions to produce interesting colour legends for visualising ternary probabilistic forecasts (see Appendix D). Theo Economou, Chris Ferro and Renato Vitolo have kindly helped with various stages of the thesis preparation. I am also grateful to the EMPS helpdesk for technical support.

My gratitude goes out to Elly Martin for PhD companionship and to my dear parents Virginia and Paul Lowe for their infinite support and efficient proof reading services. Finally, I am eternally grateful to André Costa for his loving support and the opportunity to practice Portuguese, which was a valuable contribution to this project.

Contents

1 Introduction **16**

1.1 Motivation 16

1.2 Research questions 21

1.3 Thesis plan 21

2 Background **23**

2.1 Introduction 23

2.2 Early warning systems for climate-sensitive diseases 23

2.2.1 Process-based biological models of infectious diseases 24

2.2.2 Empirical models of infectious diseases 27

2.2.3 Seasonal climate forecasts 30

2.2.4 El Niño-Southern Oscillation and infectious diseases 32

2.2.5 Existing early warning systems for climate-sensitive diseases 34

2.2.6 Candidate climate-sensitive diseases 36

2.3 Dengue 39

2.3.1 Transmission cycle 40

2.3.2 Previous climate-dengue studies 42

2.3.3 Dengue in Brazil 46

2.3.4 Surveillance and control 47

2.3.5 Climate and geography of Brazil 48

2.3.6 Climate variations in Brazil 50

2.4 Summary 51

3 Exploratory data analysis **53**

3.1 Introduction 53

3.2 Dengue data 53

3.2.1 Limitations of dengue data 56

3.2.2 Dengue incidence rate 59

3.2.3 Standardised morbidity ratios 61

3.3 Cartographic and demographic data 62

3.4 Climate data 64

3.4.1 Precipitation and temperature gridded datasets 65

3.4.2 Comparing gridded datasets to microregion data 66

3.4.3 Dengue and gridded climate 68

3.4.4 Dengue and ENSO 71

3.5 Summary 79

4 Model framework **80**

4.1 Introduction 80

4.2 Generalised linear model framework 80

4.2.1 Poisson model 82

4.2.2 Negative binomial model 83

4.3 Overdispersion 85

4.4 Goodness-of-fit 85

4.5 Choice of distribution . . . 87

4.5.1 Residual analysis . . . 89

4.6 Selection of covariates . . . 92

4.7 Robustness of ENSO effect on dengue . . . 108

4.7.1 Regional model framework . . . 108

4.7.2 Local variations in optimal time lag . . . 109

4.7.3 Influence and leverage . . . 110

4.7.4 Peak months . . . 113

4.7.5 Summary . . . 113

4.8 Conclusion . . . 114

5 Extension to a Bayesian hierarchical model framework 116

5.1 Introduction . . . 116

5.2 Generalised linear mixed model framework . . . 118

5.3 Fixed effects . . . 118

5.4 Random effects . . . 119

5.4.1 Spatially unstructured random effects . . . 119

5.4.2 Spatially structured random effects . . . 120

5.4.3 Combination of unstructured and structured random effects . . . 120

5.4.4 Temporally autocorrelated random effects . . . 121

5.5 Selection of random effects . . . 122

5.6 Model implementation . . . 124

5.6.1 Convergence of Markov chains . . . 125

5.6.2 Goodness-of-fit . . . 126

5.6.3 Inference for climate covariates . . . 130

5.7 Comparison of fixed and mixed effects model . . . 132

5.8 Climate contribution to dengue relative risk . . . 138

5.8.1 Response to ENSO . . . 138

5.8.2 Response to local climate for different ENSO scenarios . . . 142

5.9 Conclusion . . . 143

6 Towards a dengue early warning system for South East Brazil 144

6.1 Introduction . . . 144

6.2 Mathematical model based on current practice . . . 145

6.3 Posterior predictive distributions . . . 148

6.4 Visualising ternary probabilistic forecasts . . . 152

6.5 Evaluation of dengue forecasting systems . . . 157

6.6 Combining GLMM with current practice . . . 163

6.7 Conclusion . . . 168

7 Conclusions 171

7.1 Summary of main findings . . . 171

7.2 Applying model framework to other regions in Brazil . . . 173

7.3 Extending lead-time by using climate forecasts . . . 174

7.4 Considerations for operational early warning systems . . . 176

7.5 Summary . . . 178

A An algorithm for fitting generalised linear models 179

B Bayesian framework and MCMC 181

B.1 Bayesian hierarchical modelling . . . 181

B.2 Estimation by Markov Chain Monte Carlo (MCMC) . . . 182

B.3 Convergence of Markov chains 183

B.4 Deviance information criterion 184

B.5 Posterior predictive distributions 185

C WinBUGS code **186**

D Visualisation of ternary probabilistic forecasts **190**

D.1 Three category probabilistic forecasts 190

D.2 Barycentric coordinates 191

D.3 Conventional practice in climate science 192

D.4 Comparing forecasts with the climatology 193

D.5 A new colour scheme for ternary forecasts 195

Glossary of Notation **198**

Glossary of Acronyms **203**

List of Figures

1.1 Progression of types of information that can be used to indicate an impending disease epidemic . . . 17

1.2 Countries or areas at risk of dengue in 2009 . . . 19

2.1 *Aedes aegypti* infestation before, during and after concentrated eradication efforts in the Americas . . . 39

2.2 Scheme of transmission of dengue from one host to another via the vector . . . 41

2.3 Climate and biomes of Brazil . . . 49

3.1 Dengue notification form . . . 55

3.2 Spatial and temporal distribution of dengue counts in Brazil 2001-2009 . . . 56

3.3 Spatial and temporal distribution of Brazilian population 2001-2009 . . . 60

3.4 Spatial and temporal distribution of DIR in Brazil 2001-2009 . . . 60

3.5 Urban population and sanitation indicators . . . 62

3.6 Spatial distribution of altitude, urban population and geographic zones . . . 63

3.7 Altitude, urban population, geographic zone and their relation to DIR . . . 64

3.8 Annual cycle of DIR for eight geographic zones of Brazil . . . 65

3.9 DJF precipitation rate and temperature climatology in South America 2001-2009 . . . 67

3.10 DJF precipitation anomalies, South America 2000-2009 . . . 68

3.11 DJF temperature anomalies, South America 2000-2009 . . . 69

3.12 Location of microregions with respect to 2.5° × 2.5° climate grid 70

3.13 Annual cycle of DIR, precipitation and temperature in the South East Atlantic Rainforest and Caatinga zones 71

3.14 Relationship between DIR and precipitation/temperature in South East Atlantic Rainforest and Caatinga zones 72

3.15 Oceanic Niño Index January 2001 - December 2009 72

3.16 Correlation maps of DJF precipitation/temperature and lagged ONI 75

3.17 Correlation maps of DJF precipitation/temperature and lagged ONI 76

3.18 Relationship between ONI and precipitation/temperature in South East Atlantic Rainforest and Caatinga zones 77

3.19 Relationship between ONI and DIR in South East Atlantic Rainforest and Caatinga zones 77

3.20 Time series of DIR, precipitation and temperature anomalies and ONI from 2001-2009 for South East Atlantic Rainforest and Caatinga zones 78

4.1 Mean and variance of dengue counts 83

4.2 Variance in dengue counts for Poisson and negative binomial GLM 89

4.3 Kernel density estimates for observed and estimated dengue counts using Poisson and negative binomial GLM 90

4.4 Deviance residuals from Poisson and negative binomial GLM in relation to theoretical quantiles for a Gaussian error distribution 91

4.5 Deviance residuals against fitted values for Poisson and negative binomial GLM . . 91

4.6 Schematic to show time lags between dengue month of interest, precipitation, temperature and ONI 94

4.7 Parameter estimates and confidence intervals for month and zone factors 98

4.8 Scatter plot and time series of observed and model fit DIR for all months and microregions in Brazil January 2001 - December 2009 99

4.9 Scatter plot of observed and model fit DIR for Brazilian zones 100

4.10 Time series of observed and model fit DIR for Brazilian zones 101

4.11 Spatial distribution of observed and model fit DIR for FMA 2001-2009 104

4.12 Time series of observed and model DIR for FMA 2001-2009 105

4.13 Scatter plot of observed and model fit DIR for South East and North East regions 105

4.14 Time series of observed and model fit FMA DIR for South East and North East regions 106

4.15 Time series of observed and model fit FMA DIR for Rio de Janeiro and Salvador da Bahia microregions 107

4.16 Change in AIC with increasing time lag between dengue relative risk and ONI . . 110

4.17 Time series of observed and GLM fit DIR from global model and South East simplified model FMA 2001-2009 111

4.18 Leverage plots 112

5.1 Trace plot of log posterior distribution for 1000 samples 125

5.2 Potential scale reduction factor for each parameter 126

5.3 Spatial distribution of posterior mean spatially unstructured random effect $\hat{\Phi}_s$, estimated in model M2 and spatially structured random effect $\hat{\Upsilon}_s$, estimated in model M3 127

5.4 Spatial distribution of posterior mean spatially unstructured $\hat{\phi}_s$, spatially structured $\hat{v}_s$ random effects and their combined effect estimated together in model M4 128

5.5 Autocorrelation in lagged estimated deviance residuals 129

5.6 Parameter estimates and confidence intervals for fixed and autocorrelated month factor 130

5.7 Kernel density estimates for the marginal posterior distributions for the parameters associated with climate variables 131

5.8 Scatter plots and time series of observed and model fit DIR using GLM and GLMM, 2001-2009 132

5.9 Scatter plots of observed and model model fit DIR using GLM and GLMM for FMA season 2001-2009 133

5.10 Spatial distribution of observed and model fit DIR for FMA 2001-2009 135

5.11 Time series of observed and model fit DIR for South East Brazil (region level) and Rio de Janeiro (microregion level), FMA 2001-2009 136

5.12 Spatial distribution of observed, GLM and GLMM fit DIR for FMA 2008 (epidemic year) and 2005 (non-epidemic year) 137

5.13 Time series of observed and model fit DIR for GLMMs with different climate covariate combinations for FMA 2001-2009 139

5.14 Climate contribution to dengue relative risk map, FMA 2001-2009 141

5.15 Climate contribution to dengue relative risk for different ENSO scenarios 143

6.1 Time series of observed, GLMM and ARM fitted DIR 2001-2009 146

6.2 Spatial distribution of observed, GLMM and ARM fitted DIR for FMA 2002-2009 147

6.3 Posterior and posterior predictive probability densities of dengue cases for Rio de Janeiro, March 2008 149

6.4 Location of five selected microregions 150

6.5 Time series of observed, posterior mean and credible interval for DIR 2008-2009 for five selected microregions 151

6.6 Symmetric and non-symmetric category boundaries of the observed distribution of DIR in South East Brazil, FMA 2001-2007 154

6.7 Probabilistic forecasts using novel visualisation technique for FMA 2008 and 2009 156

6.8 ROC curve schematic 160

6.9 ROC curve for binary event of DIR exceeding epidemic threshold (300 cases per 100,000 inhabitants) in FMA 2008 and 2009 using GLMM 161

6.10 Posterior predictive distribution for peak dengue season 2008 and 2009 for five selected microregions 162

6.11 ROC curve for binary event of DIR exceeding epidemic threshold (300 cases per 100,000 inhabitants) in FMA 2008 and 2009 using ARM 163

6.12 Time series of observed, GLMM, ARM and combined GLMM fit DIR 2001-2009 . 165

6.13 Probabilistic forecasts using novel visualisation technique for FMA 2008 and 2009 using combined GLMM 166

6.14 ROC curve for binary event of DIR exceeding epidemic threshold (300 cases per 100,000 inhabitants) in FMA 2008 and 2009 using combined GLMM 168

7.1 Observed and GLMM fit DIR for North East Brazil, MAM season 2002 174

7.2 Schematic to show time lags between dengue month of interest and forecast issue month 176

7.3 Verification skill map of EUROBRISA forecasting system 177

D.1 Ternary phase diagrams for three category probabilistic forecasts 192

D.2 EUROBRISA ternary forecast visualised with a categorical colour scheme 194

D.3 HSV colour cone 196

D.4 Assigning colours to ternary probabilistic forecasts 197

List of Tables

2.1 Candidate diseases for climate-based early warning systems 38

3.1 Notified dengue cases and dengue incidence rate for main regions of Brazil January 2001 - December 2009 56

3.2 Source and original resolution of datasets. 58

4.1 Test results and information criteria to compare Poisson and negative binomial GLM 88

4.2 Test results and information criteria to compare models of increasing complexity . 95

4.3 Parameter estimates for climate covariates in eight zones of Brazil 96

4.4 Summary of parameter estimates for ONI and AIC at different time lags 110

4.5 Coefficient estimates for ONI by deletion of points for South East Brazil 113

5.1 Deviance results for fixed and mixed effects models 126

5.2 Parameter estimates and convergence diagnostic for climate covariates 131

5.3 Deviance results for GLMMs with different combinations of climate covariates . . . 139

6.1 Deviance results for GLMM and ARM 145

6.2 Four possible outcomes for categorical forecasts of a binary event 158

6.3 Results for epidemic prediction for the 160 microregions FMA 2008 using GLMM . 159

6.4 Results for epidemic prediction for the 160 microregions FMA 2009 using GLMM . 159

6.5 Results for GLMM, ARM and combined GLMM 165

6.6 Results for epidemic prediction for the 160 microregions FMA 2008 using combined GLMM 167

6.7 Results for epidemic prediction for the 160 microregions FMA 2009 using combined GLMM 167

6.8 Area under ROC curve for GLMM, ARM and combined GLMM 168

Chapter 1

Introduction

1.1 Motivation

The early identification of an epidemic of infectious disease is an important first step towards implementing effective interventions to control the disease and reducing mortality and morbidity in human populations (Kuhn et al., 2005). However, an epidemic is often under way before the authorities are notified and control measures are put in place. The transmission of many infectious diseases is often influenced by weather and climate variability, particularly for those spread by arthropod vectors such as malaria, dengue and the plague (Gage et al., 2008). Some vector-borne diseases demonstrate seasonal patterns and display inter-annual variability which can partly be explained by meteorological factors (Kovats et al., 2003). Therefore, climate information could potentially be valuable in early warning systems for epidemic-prone diseases.

The goal of a climate-based epidemic early warning system is to provide public health decision makers and the general public with as much advance notice as possible about the likelihood of a disease outbreak in a particular location, to allow the implementation of timely preventative measures (Burke et al., 2001). Such early warning systems require statistical and/or biological models which capture the impact of climate variables on disease transmission. Due to time lags involved in the climate-disease transmission system, lagged observed climate variables could provide some predictive lead for forecasting disease epidemics. This lead time could be extended by using forecasts of the climate in disease prediction models. However, the inherent dilemma of an early warning

system is that more lead time means less predictive certainty (see Fig 1.1). In most cases a 'surveillance and response approach' is adopted to detect impending epidemics by monitoring the appearance of early cases of the disease in a population. Although this provides relatively high predictive certainty, it leaves public health decision makers with little advance notice to implement preventative measures. Alternatively, precursory climate observations and seasonal climate forecasts could potentially be used to predict climatic conditions suitable for pathogen development and disease transmission in order to permit early efforts to minimize the spread of the disease.

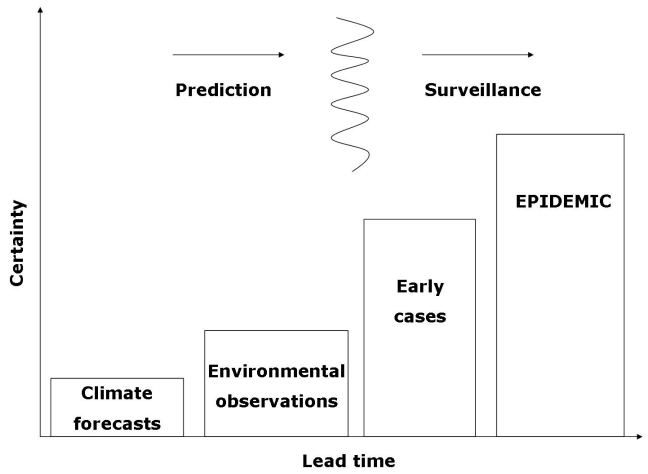


Figure 1.1: Progression of types of information that can be used to indicate an impending disease epidemic. *After* Burke et al. (2001).

Recent epidemiological studies have demonstrated statistically significant associations between climate variations and various infectious diseases (for a review see Kelly-Hope and Thomson, 2008), and have highlighted the potential for developing climate-based early warning systems (e.g. Thomson et al., 2006). However, to establish how much variation in disease risk can be attributed to climatic factors, non-climatic confounding factors must also be carefully considered in the model parameterisation to avoid bias in estimating climate-disease associations. In addition, previous models are often of limited value for public health decision making due to the lack of spatial resolution and/or the capability to predict future disease risk (e.g. several months ahead).

In this thesis, a statistical modelling framework is proposed to model disease risk using both climate and non-climate information. The framework is developed in the context of

dengue fever. The World Health Organization (WHO) has identified dengue as one of the important climate-sensitive diseases for which early warning systems should be improved (Kuhn et al., 2005). Research recommendations involve quantifying the role that climate plays in the transmission of the disease and constructing and testing predictive models. Dengue fever and its more severe form (dengue hemorrhagic fever) is one of the most important emerging tropical diseases at the beginning of the 21st century in terms of morbidity and mortality (Gubler, 2002b, Guzman and Kouri, 2003). Dengue is an acute viral disease characterised by fever, headache, severe muscle and joint pains (hence the nickname break-bone fever), rash, nausea, and vomiting (Rigau-Pérez et al., 1998). Most dengue infections do not result in death, but a small portion develop into the more serious and potentially deadly illness dengue hemorrhagic fever/dengue shock syndrome. This is characterized by spontaneous hemorrhage, increased permeability of the blood vessels and circulatory failure, leading to shock. Fatality rates in untreated dengue hemorrhagic fever/dengue shock syndrome can be as high as 50% (Reiter, 2001).

Dengue viruses are transmitted by the bite of infected *Aedes* females, in particular *Aedes aegypti*, an urban mosquito with widespread distribution in tropical cities (Hayden et al., 2010). Dengue transmission is influenced by many factors, including climate, which influences mosquito biology and interactions between the mosquito vector and dengue virus (Kuno, 1995; Scott et al., 2000; Sanchez et al., 2006). Dengue is endemic in many tropical and subtropical countries. However, epidemic dengue transmission displays a seasonal pattern in response to temperature and rainfall variability (Johansson et al., 2009a). There have been recent concerns of a worldwide spread of dengue fever, as a result of climate change, that could favour an expansion of the transmission area (Epstein, 2001; Hales et al., 2002). Global incidence of dengue has grown dramatically in recent decades and about two fifths of the world's population are now at risk (WHO[1], 2009, see Fig. 1.2[2]), with an estimated 50 million dengue infections worldwide every year. Several studies have reported associations between spatial (e.g. Hales et al., 2002) and temporal (e.g. Hales et al., 1999; Gagnon et al., 2001; Cazelles et al., 2005) patterns of dengue and climate. However, these reported associations are not entirely consistent, possibly reflecting the complexity of climatic effects on transmission, and/or the presence of confounding factors such as population immunity.

[1] http://www.who.int/mediacentre/factsheets/fs117/en/index.html, [accessed 15 May 2010]

[2] http://gamapserver.who.int/mapLibrary/Files/Maps/Global_DengueTransmission_ITHRiskMap.png

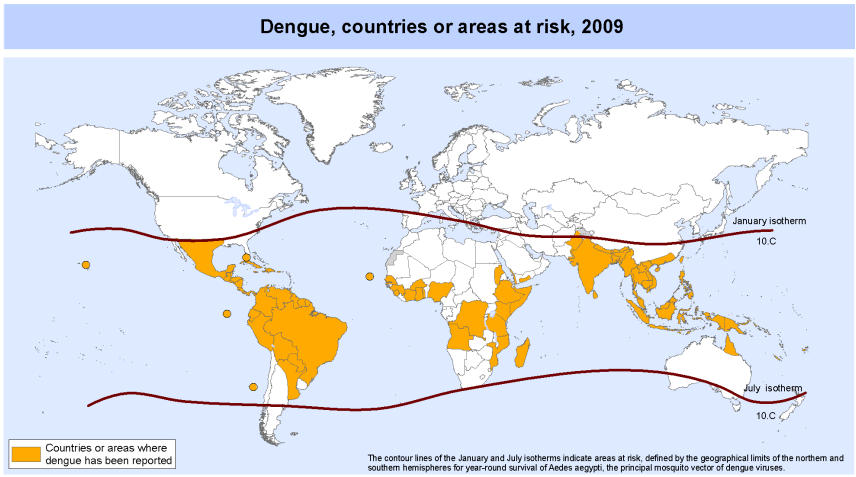


Figure 1.2: Countries or areas at risk of dengue in 2009. *Source:* WHO (2010)[2].

Brazil is used as a case study throughout the thesis to assess the potential to provide early warnings of future increased and geographically specific risk of dengue. In the 21st century, Brazil became the country with the most reported cases of dengue fever in the world, with more than three million cases reported from 2000 to 2005 (Teixeira et al., 2009). This represented 78% of all cases reported in the Americas and 61% of all cases reported to the WHO. Luz et al. (2009) assessed the dengue burden in Brazil using the non-monetary index Disability Adjusted Life Years (DALYs) (Murray and Lopez, 1994; Murray, 1994). DALYs account for the mortality, or the time lost due to premature death, and the morbidity, or the time lived with disability, imposed by a disease or health condition. From 1986 through 2006, a mean value of 56, 47 and 22 DALYs per million individuals annually were lost to dengue in the city of Rio de Janeiro, in the state of Rio de Janeiro and in Brazil, respectively. The authors state that the dengue burden in Brazil is of the same order of magnitude as the dengue burden in countries that have, for decades, severely suffered from dengue, such as Puerto Rico and Thailand. Brazil has some of the worlds best laboratory-based surveillance capabilities for dengue/dengue haemorrhagic fever (Gubler, 2002a). However, this surveillance system is not routinely used as an early warning system to predict epidemics. Therefore, Brazil serves as an excellent 'test bed' for which to develop a climate-based early warning system for dengue epidemics.

In Brazil, the greatest incidence of cases occurs from January to May when the climate

is warmest and most humid (Braga and Valle, 2007). Dengue epidemics impact heavily on national health services. There is no specific treatment for dengue, but appropriate medical care frequently saves the lives of patients with the more serious dengue haemorrhagic fever. A major epidemic occurred in Brazil in 2008, with 787,726 reported cases (January to November) including 4,137 cases of hemorrhagic fever and 448 deaths[3]. In Rio de Janeiro, military field clinics had to be installed during the 2008 outbreak to help to ease the pressure on emergency rooms packed with people suffering from dengue[4].

The current dengue surveillance system in Brazil relies on observing early cases of dengue in December/January to estimate epidemic potential later in the austral summer (February - April). However, this provides neither quantitative estimates nor a long predictive lead time. The greater the lead time available for forecasting disease risk, the greater the opportunity for effective disease risk intervention, although long term predictions often involve larger errors. Myers et al. (2000) suggested that epidemic forecasting is most useful to health services when case numbers are predicted two to six months ahead. This would allow time for the allocation of resources to interventions such as preparing health care services for increased numbers of dengue patients and educating populations to eliminate mosquito breeding sites i.e. by regularly emptying water that accumulates in discarded refuse, tyres and domestic water storage containers, commonplace in urban slums/favelas found in some areas of Brazil.

Therefore, the use of climate information with time lags relevant to dengue transmission within a spatio-temporal dengue early warning system is a research area in need of exploration. There is a need to quantify the extent to which climate information helps predict variations in dengue risk and the potential usefulness of seasonal climate forecasts with lead times of one month or more within an integrated operational dengue early warning system. The Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment report states that projected trends in climate change are expected, albeit with low confidence, to increase the number of people at risk of dengue (Confalonieri et al., 2007). Therefore, such operational systems may play an important role in adaptation strategies in the event of future climate change.

[3] http://portal.saude.gov.br/portal/arquivos/pdf/boletim_dengue_janeiro_novembro.pdf, [accessed 15 May 2010].

[4] http://news.bbc.co.uk/1/hi/world/americas/7324000.stm, [accessed 15 May 2010].

1.2 Research questions

The goal of this research is to develop a well-specified statistical modelling framework capable of providing probabilistic forecasts of disease risk in both time and space. The framework can then be used to assess the viability of incorporating climate information into a dengue early warning system for Brazil. The following questions will be addressed in this thesis:

1. To what extent can spatio-temporal variations in dengue risk be accounted for by climate variations?
2. How can observed and unobserved non-climatic confounding factors be incorporated?
3. How well can the developed model predict future and geographically specific dengue epidemics?

1.3 Thesis plan

Chapter 2 provides a literature review of climate-based disease early warning systems. Dengue fever epidemiology, the factors affecting the dengue transmission cycle and a review of previous work using statistical techniques to model climate-dengue relationships is included. Dengue, current surveillance and monitoring practices and climate variability in Brazil are also discussed. Chapter 3 explores datasets specific to dengue in Brazil, and associated climate and non-climate confounding factors relevant to dengue transmission. In Chapter 4, a generalised linear model (GLM) is introduced to explain monthly dengue counts in Brazil from January 2001 - December 2009. Model selection is used to identify climate and other covariates important for prediction. Subsequently, in Chapter 5, the selected model for Brazil is refined in the context of the South East region of Brazil, where dengue predominates. The resulting spatio-temporal hierarchical model (or GLMM - generalised linear mixed model) is estimated in a Bayesian framework using Markov Chain Monte Carlo (MCMC). In Chapter 6, the predictive power of the developed model is assessed. The model is refitted to data from 2001-2007 and evaluated by comparing posterior predictive distributions to out-of-sample data 2008-2009. This

is compared to a simple model representative of current practice for dengue surveillance in Brazil. The forecasting systems are evaluated given a pre-defined epidemic threshold and varying probabilistic decision thresholds. The potential benefit of combining the developed forecasting system with current practice is then considered. A novel visualisation technique is presented and used to communicate ternary probabilistic forecasts for dengue risk using the proposed forecasting systems. The final chapter summarises the findings of this thesis and recommends future research directions.

Chapter 2

Background

2.1 Introduction

The aim of the first part of this chapter is to give a brief overview of climate-based disease early warning systems and some of the modelling techniques that have been used to predict climate-sensitive disease risk. The second part of the chapter focuses on dengue fever; its epidemiology, the role of climate in the dengue transmission cycle and previous studies linking climate to dengue worldwide. Dengue, surveillance and monitoring and climate variations in Brazil are then discussed. This review will help to identify and highlight the knowledge needed to develop a successful predictive model for dengue risk based on climate information.

2.2 Early warning systems for climate-sensitive diseases

It is commonly accepted that climate plays a role in the transmission of many infectious diseases, some of which are among the most important causes of mortality and morbidity in developing countries (Kuhn et al., 2005). Climate variability can affect infectious disease transmission, both in terms of spatial and seasonal distribution, inter-annual variability and epidemic potential. Certain infectious diseases are more sensitive to climate variability than others. The effect that climate variability has on infectious disease is determined largely by the unique transmission cycle of each pathogen (McMichael et al.,

2003). Transmission cycles that require a vector or non-human host (e.g. mosquitoes) are more susceptible to external environmental factors including temperature, precipitation and humidity. Diseases transmitted by mosquitoes, such as malaria and dengue fever, are particularly sensitive to weather conditions. Rainfall can affect the availability of mosquito breeding, developmental and resting sites (Hunter, 2003). Temperature influences the rate of development of immature stages and adult survival rate (Rueda et al., 1990), biting frequency (Githeko et al., 2000) and the extrinsic incubation periods (the period between infection of the vector and the vector's ability to infect the next susceptible host) of disease agents (Watts et al., 1987).

The effects of climate variability and change on the epidemiology of mosquito-borne viral diseases are not easily predictable (Gage et al., 2008). Cursory considerations might conclude that increases in temperature and precipitation will produce increased incidence of mosquito-transmitted diseases. However, the ecologic determinants of these diseases interact in complex ways. The incidence of dengue and chikungunya fever, both transmitted by *Aedes aegypti* mosquitoes, sometimes increases during dry seasons because of increased peri-domestic water storage (Pontes et al., 2000; Chretien et al., 2007). Extremely high temperatures can increase mosquito mortality and in the case of malaria, parasite development cannot occur above temperatures of 33-39°C (Gage et al., 2008). This indicates that increasing temperatures may restrict mosquito-transmitted diseases in some geographic regions. Rainfall can promote transmission by creating ground pools and other breeding sites, but heavy rains can have a flushing effect, cleansing such sites of their mosquitoes (Reiter, 2001). Reiter et al. (2003) found that the difference in dengue incidence between two contiguous cities that straddle the US-Mexico border was likely due to the use of air-conditioning and human behaviour rather than climatic factors, since the climate was identical and favourable to dengue transmission for both cities. When attempting to model the effect of climate on such diseases, careful consideration of the interaction of ecologic variables with human behaviour and the urban environment is important.

2.2.1 Process-based biological models of infectious diseases

Mathematical modelling is increasingly being applied to interpret and predict the future incidence and control of infectious diseases. While statistical models are driven by data

and use empirical relationships between the disease and climate variables, process-based models use differential equations to represent the dynamical evolution of the disease lifecycle and incorporate climate variables as parameters (Jones, 2007). Most process-based models consider the likelihood that the habitat for vectors will be suitable for perennial, seasonal or epidemic transmission of the disease (Lafferty, 2009). Process-based and rules-based models of malaria risk have been created for mapping and climate change applications (e.g. Martens et al., 1999; Craig et al., 1999; Tanser et al., 2003). For example, Craig et al. (1999) describe a fuzzy logic rules-based model to define the crude distribution of malaria transmission in Africa, based upon biological constraint of climate on parasite and vector development.

More sophisticated mathematical models have been developed for malaria epidemic prediction (e.g. Githeko and Ndegwa, 2001; Hoshen and Morse, 2004; Worrall et al., 2007). Hoshen and Morse (2004) formulated a dynamic process-based malaria model to represent the climate-driven biological mechanisms associated with transmission of the disease. Modelled cases were compared to reported cases for Hwange, Zimbabwe 1995-1998. The model was able to capture both the seasonality and the inter-annual variability of infection at the test site in Zimbabwe. This model was used in a subsequent study to validate a seasonal climate forecast system using a three-tier hierarchical approach (Morse et al., 2005). Tier-1 refers to the verification of relevant climate variables against observations of those variables. Morse et al. (2005) performed a tier-2 validation using probabilistic hindcasts of simulated malaria prevalence driven by rainfall and temperature forecasts from the DEMETER multi-model ensemble prediction system (Palmer et al., 2004) and malaria prevalence estimates driven with ERA-40 gridded reanalysis data. The validation was carried out for four model grid points, at 2.5° resolution, in southern Africa for the period 1987-2001. Results showed that the DEMETER-driven malaria prevalence hindcasts were skillful when compared against malaria modelled using ERA-40 reanalysis data (tier-2 validation) for the one-month lead seasonal predictions. When the analysis focused on the event 'prevalence above the median', malaria DEMETER driven hindcasts were also found to be tier-2 skillful for the period covering the seasonal malaria peak with a 4-6 month forecast window. In an ideal situation, the integrated probabilistic forecasting system for malaria would be validated against reported malaria data (tier-3 validation). However, for most parts of Africa sufficiently high quality clinical data does not exist. In a subsequent study, Jones and Morse (2010) validated the forecasting sys-

tem against reported malaria data for Botswana (tier-3 approach). 20 years of malaria data published by Thomson et al. (2005) were used to create three time-series of binary events: low malaria years (observed malaria anomalies below the lower tercile), above average malaria years (observed malaria anomalies above the median) and high malaria years (observed malaria anomalies above the upper tercile). Results showed that the tier-3 skill of DEMETER-driven malaria forecasts (validated against reported malaria data) was better for predicting malaria in the lower tercile rather than the upper tercile.

Focks et al. (1995) developed stochastic simulation models that describe the daily dynamics of dengue virus transmission in the urban environment. These models take into account the majority of factors known to influence dengue epidemiology. The first model, the container-inhabiting mosquito simulation model (CIMSiM), a weather-driven dynamic life-table model of container-inhabiting mosquitoes such as *Aedes aegypti*, provides inputs to the transmission model, the dengue simulation model (DENSiM). Together, CIMSiM and DENSiM incorporate virtually all of the commonly recognised factors influencing the dynamics of dengue viruses in the urban setting. Model simulations have been shown to provide a good description of temporal variations in mosquito population dynamics in Bangkok and New Orleans, and of the seasonal pattern of transmission during an epidemic in Honduras (Focks et al., 1993a,b, 1995). This model represents a full biological approach to an early warning system, and requires specific information on a range of parameters such as mosquito breeding, population density, virus serotypes and vertebrate hosts. However, from a public health decision making point of view, such monitoring may be too costly and time consuming for use in developing countries (Kuhn et al., 2005).

At a global scale, Patz et al. (1998) use General Circulation Models (GCMs) to estimate the potential contribution of climate change to enhancing the capacity of vectors to transmit dengue. Dengue specific variables were derived from the relationships used in CIMSiM and DENSiM. GCMs projected a temperature-related increase in potential seasonal transmission in five selected cities (Bangkok, San Juan, Mexico City, Athens and Philadelphia), as well as an increase in global epidemic potential, with the largest area change occurring in temperate regions. The authors suggest that such global climate scenario-based analyses should be integrated with local demographic and environmental factors to guide comprehensive and long-term preventive health measures.

As process-based models are based on underlying physical and biological processes, they can arguably be applied to regions where reliable data is lacking, or to predict future scenarios. However, such models are limited by understanding of the biological mechanisms involved, the omission of significant aspects of the vector or parasite lifecycle (due to the lack of information in the literature) and by the availability of data for model input and model validation (Jones and Morse, 2010).

2.2.2 Empirical models of infectious diseases

Statistical disease models based on past empirical data differ from process-based models (see Section 2.2.1) in that they do not assume a functional relationship between vital rates of vectors and environmental variables, allowing a wider range of explanatory variables to be included in the model parameterisation (Lafferty, 2009). The availability of large scale spatial datasets allow complex analyses at global scales. Despite the increasing number of statistical models to predict infection risk for a range of diseases, the assessment of their spatial limits, predictive performance and practical application are not widely undertaken (Brooker et al., 2002). There are several limitations to the statistical models that have been developed in the climate and health literature. These limitations are discussed below.

Models to predict vector-borne disease often include an autoregressive time series component (e.g. Zhou et al., 2004; Gomez-Elipe et al., 2007; Tipayamongkholgul et al., 2009), based on the idea that the current value of the time series can be explained as a function of past values. However, previous studies do not always quantify the contribution of an autoregressive lagged disease term compared to climate covariates to the variance explained by such a model. For example, Tipayamongkholgul et al. (2009) model dengue incidence in Thailand using 1 month lagged dengue incidence and climate variables. The reported R^2 values suggest that the incorporation of both El Niño and local climate data into province specific models explains a large proportion of the variance in dengue incidence. However, the authors do not investigate the influence of the local climate alone, which may account for very little of the variance explained. Further, autoregressive terms with only one month lag offer little, if any, advance warning of an impending epidemic. In practice, the collation of such data may not be feasible in advance of the time period for which the forecast is valid.

Some models assume a probability distribution that might not be suitable for the disease variable in question. It is important to recognise whether the response variable takes real values (e.g. disease incidence) or integer values (e.g. counts of cases). When modelling disease incidence, a Gaussian probability distribution may be appropriate. However, when modelling counts, a suitable probability distribution should be deployed, such as the Poisson or negative binomial. For example, Zhou et al. (2004) modelled the effects of autoregression, seasonality and climate variability on the number of malaria inpatients in seven sites in the East African highlands, using Gaussian linear regression. However, a probability distribution for count data might have been more appropriate.

A common issue when modelling counts of cases is the presence of extra-Poisson variation, or overdispersion (variance greater than the mean) in the data, given a model where the variance is a function of the mean (see Chapter 4 for more details). Overdispersion requires attention in model fitting (Crawley, 2002; Venables and Ripley, 2002). A well-established method is to use a negative binomial model, which has a scale parameter that can be used to account for overdispersion. This scale parameter can be used to adjust the variance independently of the mean (Venables and Ripley, 2002). Mantilla et al. (2009) used both a Poisson and a negative binomial regression model to explain annual counts of malaria cases in Colombia, using annual indices of the El Niño Southern Oscillation (see Section 2.2.4). They found that the negative binomial model was better as the malaria cases were overdispersed compared to a Poisson model.

Overdispersion or spatial correlation due to unobserved confounders will usually not be captured by simple covariate models and often it is appropriate to include some additional term or terms in a model which can capture such effects (Lawson, 2008). In most cases, data on confounding factors such as population immunity and quality of health care services or interventions are not available. The inclusion of random effects in the model framework can help to account for such unknown or unobserved confounding factors in the disease systems.

Defining the spatial distribution of a disease within a country or region is a fundamental step to understanding its epidemiology (Kelly-Hope and Thomson, 2008). It allows public health decision makers to identify zones susceptible to epidemics and to identify vulnerable groups at risk. Spatial maps can allow the comparison of different diseases, the analysis of temporal/inter-annual variations across the space and the influence of

climate and other confounding factors in the spatial heterogeneity of the disease. In the United States, fine-resolution human plague risk models based on landscape and ecological features have been constructed (Eisen et al., 2007a,b). A similar spatial risk model has been developed to define areas of plague risk in the West Nile Region of Uganda (Winters et al., 2009). Previous studies (Parmenter et al., 1999; Enscore et al., 2002; Stapp et al., 2004; Snall et al., 2008) have indicated that precipitation and other climatic variables are likely to influence the frequency of human plague, in several geographic locations, by affecting the spread of plague among the rodents and fleas that act as sources of infection for humans (Gage et al., 2008). Therefore incorporating spatio-temporal climate information within these models could be beneficial.

Omumbo et al. (2005) developed a malaria risk map for East Africa using satellite sensor-derived data including the normalized difference vegetation index (NDVI - an indicator of photosynthetic activity and a surrogate for moisture availability) and land surface temperature. The model included non-climatic data such as urbanisation and presence of water bodies. The map was further divided into two ecological zones, as the factors influencing vector distribution and abundance vary between these ecological zones.

Such mapping efforts aid the targeting of limited public health resources toward those areas at greatest risk. By further incorporating temporal information based on climatic or other temporally varying variables in these models, early warnings of geographically specific increased levels of disease may be possible. As a disease may exhibit different annual patterns in different ecological zones within the country or region of interest (Kelly-Hope and Thomson, 2008), it is important to consider such differences when modelling a disease over a large geographical area. Further, statistical models should include potentially important non-climate variables that might lead to the absence of infectious diseases in areas with climate disposed to infectious disease transmission (Lafferty, 2009).

Bayesian geostatistical approaches are increasingly used for mapping and predicting the risk of infectious diseases. Diggle et al. (2007) extended a spatial logistic regression model of *Loa loa* (a filarial worm that can adversely interfere with the treatment given for onchocerciasis) prevalence in Cameroon (Thomson et al., 2004). This issue of spatial correlation was addressed and Bayesian methods were used to quantify the uncertainty in the predictions from the new model (Diggle et al., 1998). Instead of mapping point estimates of prevalence, the probability, given the data, that a particular location did or

did not exceed a pre-defined threshold was mapped. The extended model (Diggle et al., 2007) was found to be more accurate than the earlier model (Thomson et al., 2004) as an unobserved spatial term was included in the model to capture residual spatial variation, after adjusting for elevation and NDVI (explanatory variables included in both models).

Mabaso et al. (2006) used Bayesian negative binomial models for the spatio-temporal analysis of the relationship between annual malaria incidence and selected climate covariates at a district level in Zimbabwe from 1988-1999. Spatial correlation was incorporated by assuming a conditional autoregressive (CAR) process in the random effects (see Chapter 5, section 5.4.2 for more details). Temporal random effects were also used at yearly intervals. Their model revealed a spatially varying risk pattern that was not attributable only to climate. Although the inclusion of temporal random effects results in an improved descriptive model, these effects are not useful for prediction, as the temporal randomness for future years is unknown and cannot be estimated. Lowe et al. (2010) used a spatio-temporal Bayesian hierarchical mixed effects model to predict dengue risk in Brazil using both climate and non-climate covariates. A more detailed study is presented in this thesis.

2.2.3 Seasonal climate forecasts

The success of a climate based epidemic early warning system is dependent on the ability of climate models to predict climatic conditions several months in advance. Seasonal climate forecasts can be produced using empirical regression methods (e.g. Folland et al., 2001; Coelho et al., 2004) or from ensembles of dynamical climate model predictions (Stephenson et al., 2005; Doblas-Reyes et al., 2006). Such climate models are an extension of the numerical methods used to predict the weather a few days ahead and are used to predict major climatic trends from one month to a few seasons ahead. Predictions of climate system evolution on seasonal timescales are subject mainly to two sources of uncertainty: uncertainty in initial conditions and model uncertainty (Doblas-Reyes et al., 2005). To address uncertainty in initial conditions, forecast models are run many times from slightly different initial conditions. The resulting ensemble of forecasts can be used to produce a forecast probability density function of the observed target variable (Stephenson et al., 2005). It has been shown that probability forecasts derived from an ensemble prediction system are of greater benefit than a deterministic forecast produced

by the same model and that, for many users, the probability forecasts have more value than a shorter-range deterministic forecast (Hagedorn et al., 2005). To address model uncertainty, a multi-model approach can be adopted (e.g. Palmer et al., 2004). Multi-model ensembles have shown clear superiority in terms of forecast quality over single-model ensembles (see Hagedorn et al., 2005; Stephenson et al., 2005; Coelho et al., 2006).

The atmosphere responds to tropical sea surface temperature (SST) anomalies such as those which occur during El Niño Southern Oscillation (ENSO) events. This makes seasonal climate forecasts of temperature and precipitation possible for specific seasons and locations, particularly in the tropics (Goddard and Mason, 2002). For example, seasonal climate forecasts have been shown to have some skill in parts of South America (Montecinos et al., 2000; Folland et al., 2001; Coelho et al., 2006; Goddard and Mason, 2002) including areas of Brazil. Seasonal climate forecasts have the potential to contribute as one component in integrated early warning systems for climate-sensitive diseases such as malaria and dengue (Connor and Mantilla, 2008). The successful implementation of such a system depends on close collaboration between public health specialists, climate scientists and mathematical/statistical modellers. Climate forecast models are currently run at coarse spatial resolutions of the order of a couple of degrees of latitude and longitude (Coelho et al., 2006). However, end user application models often require climate information at a finer spatial and/or temporal resolution. Therefore, downscaling of spatially and temporally coarse model output to finer scales in areas of interest is often required. Downscaling can be performed using dynamical (e.g. Misra et al., 2003) or statistical methods (e.g. Feddersen and Andersen, 2005; Stephenson et al., 2005; Coelho et al., 2006). Statistical downscaling techniques depend on the availability of reliable past observational data. Availability of observational data to develop statistical/empirical downscaling methods is a major problem for tropical regions[1]. This can be overcome using dynamical techniques where regional climate models are run at a high spatial resolution. However, this method is computationally expensive and there is no means of checking performance with observed data, i.e. for validation and model calibration purposes.

Seasonal climate forecasts are often issued as probabilities for tercile categories (e.g. below normal, normal, above normal), which can be difficult to incorporate within the

[1] http://www.ecmwf.int/research/EU_projects/ENSEMBLES/documents/20050526_s2ddownscaling.pdf [assessed 19 August 2010]

context of specific health early warning systems. The format of forecast data should ideally be tailored to suit specific disease control problems (Connor and Mantilla, 2008). For research purposes, hindcast data (i.e. retrospective forecasts made for a historical period in pseudo-operational mode) from seasonal climate forecasting centres such as the UK Met Office and the European Centre for Medium Range Forecasts (ECMWF) can be requested. However, real-time seasonal climate forecast data are not always easily obtainable or freely available to public health decision makers. Observed climate data are readily available via the internet (e.g. IRI/LDEO Climate Data Library[2], NOAA/ESRL/PSD gridded climate datasets[3]). Such climate data can provide public health services with useful indicators of changes in epidemic risk. Although lead-times are shorter than for seasonal climate forecasts (1-3 months), they are generally more certain as they are based on observations.

2.2.4 El Niño-Southern Oscillation and infectious diseases

The El Niño Southern Oscillation (ENSO) is a coupled oceanic-atmospheric phenomenon, characterised by sustained fluctuations between unusually warm (El Niño) and cold (La Niña) conditions in the tropical Pacific Ocean. El Niño and La Niña events typically recur every 2 to 7 years and develop in association with large-scale oscillations in an atmospheric pressure pattern spanning the tropical Indian and Pacific Oceans, known as the Southern Oscillation (Philander, 1990; McPhaden et al., 2006). ENSO influences the inter-annual variability in weather patterns and the likelihood of regional extreme events, such as droughts and floods, across the globe (Ropelewski and Halpert, 1987; Lyon and Barnston, 2005). The effect of ENSO on climate variations in Brazil is discussed further in Section 2.3.6.

An association between ENSO and a heightened risk of certain vector-borne diseases has been identified in specific geographical areas where climate anomalies and ENSO are linked (Kovats, 2000; Kovats et al., 2003). For example, Bouma et al. (1997) found a positive association between El Niño events and malaria risk in Colombia for the period 1960-1992. A similar association was found for Venezuela for the period 1975-1990 (Bouma and Dye, 1997). Gagnon et al. (2002) analysed the relationship between ENSO

[2] http://iridl.ldeo.columbia.edu/index.html [accessed 19 August 2010]

[3] http://www.esrl.noaa.gov/psd/data/gridded/ [accessed 19 August 2010]

events and malaria epidemics in a number of tropical South American countries. A statistically significant relationship was found between El Niño and malaria epidemics in Colombia, Guyana, Peru, and Venezuela. Flooding in the dry coastal region of northern Peru, which may be induced by El Niño, coincided with high malaria incidence, while El Niño triggered drought conditions were associated with malaria epidemics in Columbia, Guyana and Venezuela. Statistically significant associations between ENSO and dengue outbreaks have been found in Thailand (Cazelles et al., 2005), South Pacific (Hales et al., 1996, 1999) and South America (Gagnon et al., 2001). Associations between ENSO and cholera were also found in Bangladesh (Pascual et al., 2000; Rodó et al., 2002) and Peru (Lama et al., 2004). Many ENSO time-series studies use aggregated national data. However, aggregation removes spatially variability in local climate. Thus, analysis at a finer sub-national geographical scale are needed to understand the complex relations between the disease and local drivers such as temperature and rainfall. The association between climate variables (temperature, rainfall) and disease should be evaluated since these variables are the principal drivers of the biological processes by which ENSO affects health (Kovats, 2000). However, only a selection of these studies report such analyses (e.g. Bouma and Dye, 1997; Pascual et al., 2000; Poveda et al., 2001). Rather than relating ENSO events to particular disease outbreaks, it may be more useful to use indices that determine ENSO events and provide a continuous indication of climate conditions (Kelly-Hope and Thomson, 2008). The Niño 3.4 index is an example of an index that is widely used to monitor the strength of El Niño and La Niña events (Barnston et al., 1997). It is defined as the anomaly in monthly SSTs over the region (120°W-170°W and 5°S- 5°N). Another commonly used index is the Southern Oscillation Index (SOI), defined as the normalised atmospheric pressure difference between Darwin in Australia and Tahiti in the South Pacific (e.g. Hales et al., 1996; Rodó et al., 2002). When developing statistical models to predict disease risk, a continuous monthly ENSO index is preferable to a yearly binary indication of an event occurring or not. The strength and duration an ENSO the event may have implications for the way in which regional climate is influenced, which in turn can influence disease variations.

2.2.5 Existing early warning systems for climate-sensitive diseases

The purpose of epidemic early warning systems is to forecast when and where an epidemic is likely to occur. Predicting the number of disease cases that could occur or the probability of exceeding an epidemic threshold, is useful for epidemic response planning. The implementation of a timely and effective response plan and ongoing evaluation of the system and its components are vital (Ebi, 2009). To be effective, public health must move from a focus on surveillance and response to a greater emphasis on prediction and prevention. An effective epidemic early warning system should reduce vulnerability of human populations to current climate variability and other confounding factors. At the same time, the system should be designed for easy modification to take into account continuing or future climate change (Ebi, 2009).

A number of recent publications describe approaches for disease forecasting using climate data. Most models aim to predict malaria epidemics in Africa, where inter-annual climate variability drives both mosquito vector dynamics and parasite development rates (Thomson et al., 2006). As monitoring of rainfall forms the basic component of a malaria early warning system (Thomson and Connor, 2001), the International Research Institute for Climate and Society (IRI) has developed an online resource freely available to national malaria control programmes that produces maps of epidemic malaria risk based on rainfall anomalies in sub-Saharan Africa (Grover-Kopec et al., 2005). Climatic conditions are considered to be suitable for transmission when the monthly precipitation accumulation is at least 80mm, the monthly mean temperature is between 18°C and 32°C and the monthly relative humidity is at least 60% (Adjuik et al., 1998; Grover-Kopec et al., 2005).

Thomson et al. (2006) developed a system to forecast probabilities of anomalously high and low malaria incidence in Botswana based on multi-model ensemble predictions of climate. An earlier study (Thomson et al., 2005) indicated that December-February seasonally averaged monitored rainfall and SSTs for the Niño 3.4 region provided significant predictive skill for the malaria season in Botswana one month in advance of its seasonal peak (1982–2002). A quadratic relationship was assumed between rainfall and malaria incidence. In Botswana, malaria epidemics peak in March and April following the bulk of the rainy season (November - February). Thomson et al. (2006) found that replacing observed precipitation with predictions from the DEMETER project (Palmer et al.,

2004) added up to four months lead time over malaria warnings issued with observed precipitation. This forecast system has been successfully applied to the prediction of malaria risk in Botswana, where links between malaria and climate variability are well established (Connor and Mantilla, 2008).

An early warning system based on this approach has also been developed in Eritrea for operational use (Ceccato et al., 2007). Relationships were investigated between monthly clinical malaria incidence in 58 districts and monthly climate data and seasonal forecasts. Although rainfall (lagged by 2 months) accounted for a relatively high proportion of the variability in malaria anomalies, the meteorological stations did not have sufficient coverage to be widely useful. Satellite derived rainfall explained some of the variability in malaria anomalies, while NDVI anomalies explained a large portion of the variability. Eritrea has two distinct rainy seasons in different parts of the country. The seasonal forecasting skill for the June - August season was low except for the Eastern border. For the coastal October - December season, forecasting skill was good only during the 1997-1998 El Niño event. Therefore, the authors recommend that for epidemic control, shorter-range warnings based on remotely sense rainfall estimates are feasible.

Jones et al. (2007) examined the relationship between climate and malaria incidence in Kegera, Tanzania with the aim of determining whether seasonal climate forecasts may assist in predicting malaria epidemics. Multiple linear regression was performed using normalised climate variables (daily meteorological data from the nearest station, aggregated to the monthly level and DEMETER hindcasts) and malaria incidence for each of the two annual malaria seasons (1990-1999). Malaria was found to be positively correlated with rainfall during the first season (October-March). For the second season (April - September), high malaria incidence was associated with increased rainfall but also with high maximum temperature during the first rainy season. The authors found that the 1997 and 1998 epidemic years significantly influenced the fit of the models. The robustness of the statistical models was tested by excluding the two epidemic years. This resulted in excessive rainfall ceasing to be a statistically significant predictor for malaria during the first season. If DEMETER had been used operationally, temperature forecasts (but not rainfall) would have correctly predicted the 1998 epidemic. However, the seasonal forecast for 1996 would have resulted in a false alarm. In 1999, the forecast and observed temperature anomalies were positive, but near normal malaria conditions were recorded. This might be explained by the rise in immunity that likely followed

the 1997 and 1998 epidemics. The authors point out that the underlying relationship between rainfall and malaria in this location may be too complicated to be revealed using regression analysis.

A major barrier to developing disease early warning systems is the shortage of time series of good quality disease data and the lack of spatial data at the sub-national level. Thomson et al. (2005, 2006) were able to use a good quality data set for malaria in Botswana that spanned more than a decade. However, because the modelling was carried out using highly aggregated data (national/annual level) for both malaria and rainfall, it was not possible to examine the timing of onset of an epidemic or area-specific variations in malaria in relation to rainfall. Epidemics are often sudden and unexpected, and prevention and control strategies need to be accurately targeted in both time and space if they are to stand a chance of being effective (Cox and Abeku, 2007).

The tendency of a disease forecasting system to issue false alarms (issuing an epidemic warning when no epidemic is later observed) or to miss an epidemic can have serious consequences. Not only in terms of morbidity and mortality, but also in terms of economic cost and the willingness of the public to rely on subsequent warnings (Ebi, 2009). Such uncertainties should be incorporated into an early warning system. In all cases, model performance should be validated using out-of-sample data (Cox and Abeku, 2007).

2.2.6 Candidate climate-sensitive diseases

It is important to identify those diseases for which climate-informed predictions offer the greatest potential for disease control. Kuhn et al. (2004) examined the most important infectious diseases identified by Murray and Lopez (2002) in the WHO global burden of disease assessment, measured in Disability Adjusted Life Years (DALYs, see Chapter 1, page 19), for climate sensitivity. Several diseases were identified as candidates for climate-based early warning systems as a means of improving preparedness and response plans for epidemics. For diseases relevant in South America, climate was identified as an 'important factor' for determining disease epidemics of cholera and malaria and 'to play a significant role' in determining disease epidemics of dengue and St Louis encephalitis (see Kuhn et al., 2004, 2005, Table 2.2.6).

Dengue fever has been selected here for further analysis. Data for St Louis encephalitis

is not readily available, as surveillance of this disease is yet to be implemented in South America. While climate-based early warning systems have been investigated for malaria (e.g. Abeku et al., 2004; Morse et al., 2005; Ceccato et al., 2007; Thomson et al., 2006; Mabaso et al., 2006) and cholera (e.g. Pascual et al., 2000; Rodó et al., 2002; Gil et al., 2004), limited progress has been made in developing early warning systems for dengue. Research recommendations for improving dengue early warning systems (see Table 2.2.6) include quantifying the role that climate plays in the transmission of the disease and constructing and testing predictive models. These recommendations form the objectives of this thesis.

Table 2.1: Common communicable climate-sensitive diseases in South America and research recommendations for improving early warning systems. *Adapted from:* Kuhn et al. (2004) and Kuhn et al. (2005).

Disease	Transmission	Climate-epidemic link	Key variables of interest	Proposed actions
Cholera	Food- and water-borne transmission	Increases in sea and air temperatures as well as El Niño events associated with epidemics. Sanitation and human behaviour also are important.	Zooplankton abundance, SST, ENSO, human factors, socioeconomic variables	1. Operational testing of Bangladesh and Peru predictive models, 2. Maintain and improve surveillance (Africa), 3. Quantify the role of climate in Africa.
Malaria	Transmitted by the bite of female *Anopheles* mosquito	Changes in temperature and rainfall associated with epidemics. Many other locally relevant factors include vector characteristics, immunity, population movements, drug resistance etc.	Temperature, rainfall, ENSO, EIR (Entomological Inoculation Rate), vector abundance, population immunity, control activities.	1. Maintain surveillance and extend to new areas, 2. Quantify role of climate (Africa and Asia), 3. Further test predictive models, 4. Link predictions to operational decisions.
Dengue	Transmitted by the bite of female *Aedes* mosquito	High temperature, humidity and heavy rain associated with epidemic. Non-climatic factors may have more important impact.	Dengue seroprevalence, socioeconomy, virus type, human immunity, precipitation, temperature and humidity	1. Quantify role of climate, 2. Maintain and improve surveillance, 3. Simplify CIMSiM and DENSiM[a] models, 4. Construct and test predictive models.
St Louis encephalitis	Transmitted by the bite of female *Culex* and *Aedes* mosquito	High temperature and heavy rain associated with epidemic. Reservoir animal factors also are important.	Bird infections, temperature, rainfall	1. Quantify role of climate, 2. Set up surveillance in South America, 3. Construct predictive models for US.

[a]Container-inhabiting mosquito simulation model (CIMSiM) and dengue simulation model (DENSiM) (Focks et al., 1995).

2.3 Dengue

Dengue fever and its more severe form dengue hemorrhagic fever is one of the most important emerging tropical diseases at the beginning of the 21st century (Gubler, 2002b). The vector responsible for major dengue epidemics is the domestic, container breeding *Aedes aegypti* mosquito (McMichael et al., 1996). Dengue may also be transmitted by *Aedes albopictus* (Monath, 1994). Dengue occurs principally in the tropical areas of Asia, Oceania, Africa, and the Americas (McBride and Bielefeldt-Ohmann, 2000). In the 1940s an *Aedes aegypti* eradication programme was initiated by the Pan American Health Organization (PAHO) in order to prevent urban epidemics of yellow fever and other vector-borne disease in the Americas (Guzman and Kouri, 2003). Unfortunately, eradication was difficult to sustain and in the 1970s the American program was disbanded and *Aedes aegypti* re-invaded most countries in the region (see Fig 2.1). The resurgence of epidemic dengue fever and the emergence of dengue hemorrhagic fever in the last few decades have been closely tied with population growth, urbanization and air travel (Gubler, 1998; Gubler and Meltzer, 1999).

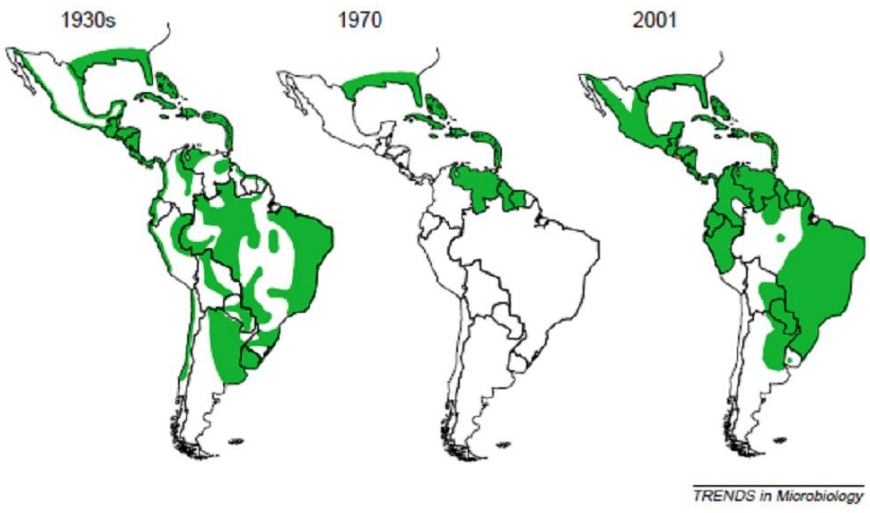


Figure 2.1: Areas with *Aedes aegypti* infestation (shaded) before, during and after concentrated eradication efforts in the Americas *Source:* Gubler (2002b).

Dengue is caused by any of four closely related dengue virus strains or serotypes (DENV-1,2,3 and 4), belonging to the family Flaviviridae (Chambers et al., 1990). Infection

with one serotype provides life-long immunity against further infection from that same serotype but no protection against the other serotypes. In fact, it has been hypothesised that sequential infections with other serotypes increases the risk of more severe manifestations including dengue hemorrhagic fever and dengue shock syndrome (Halstead, 1981). There is no specific treatment for dengue. For dengue hemorrhagic fever patients, maintenance of circulating fluid volume in the body is extremely important[4]. Despite significant progress in vaccine development (Halstead and Deen, 2002; Webster et al., 2009), there is no tested and approved vaccine to protect against dengue. Therefore, disease control and prevention have mainly focused on vector control activities and surveillance (Siqueira et al., 2005; Teixeira et al., 2009).

Although dengue is not as deadly as malaria, the sheer number of people infected and the fevers debilitating nature mean that it has an enormous economic impact. In terms of disruption to quality of life and economic productivity, the burden of dengue on some societies is comparable to that of HIV, tuberculosis or hepatitis (Clarke, 2002). However, the control, treatment and management of these diseases receive a great deal more attention and funding than dengue from international funding agencies. For example, in 1998, it is estimated that $84 million was directed at the global malaria problem, whereas the problem of dengue and dengue hemorrhagic fever received less than $5 million (Gubler and Meltzer, 1999).

2.3.1 Transmission cycle

Symptoms of infection in a human host usually begin 4 - 7 days after the mosquito bite and typically last 3 - 10 days[5] (see Fig. 2.2). In order for transmission to occur, the mosquito must feed on a person during a 5 day period when large amounts of virus are in the blood. This period usually begins just before the person becomes symptomatic. A person can be infected with a dengue virus without showing significant symptoms. However, the virus can still be transmitted to a mosquito. The virus requires a further 8-12 day extrinsic incubation period before it can be transmitted to another human[5]. The mosquito remains infected for the remainder of its life, which could be a few days or several weeks. As the blood meal stimulates oviposition by the female mosquito,

[4] http://www.who.int/mediacentre/factsheets/fs117/en/index.html, [accessed 15 May 2010]

[5] http://www.cdc.gov/dengue/epidemiology/index.html [accessed 15 May 2010]

which undergoes at least one, and often more, reproductive cycles during the extrinsic incubation period, there is an opportunity for virus to enter the egg and be passed to the next generation of mosquitoes (Monath, 1994).

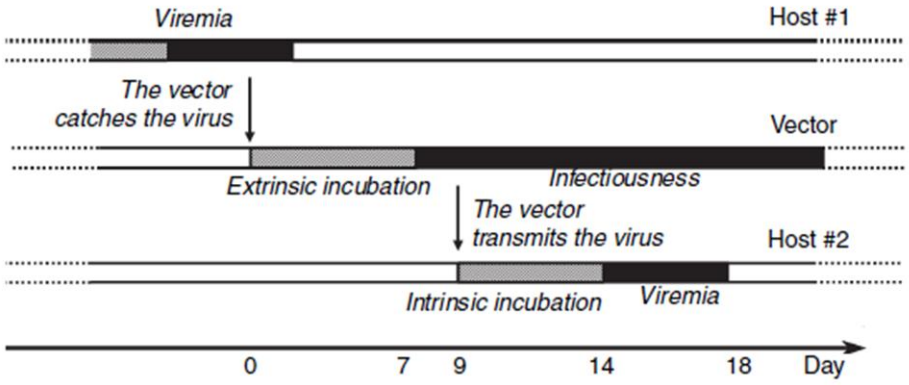


Figure 2.2: Scheme of transmission of dengue from one host to another via the vector. The vector infectiousness lasts for the rest of its life. In this example, the extrinsic incubation period lasts 7 days, the intrinsic incubation period lasts 5 days and the viraemia (the presence of virus in the blood) up to 4 days. *Source:* Favier et al. (2006).

A number of complex factors are related to dengue transmission, in particular population growth and unplanned urbanization, resulting in substandard housing, inadequate water, sewerage and waste management systems which allow mosquito reproduction (Gubler, 2002b). Dengue incidence is also influenced by mosquito population densities and survival, type and productivity of containers that hold larvae, adult flight range, human population density, virus strain, immunity for specific virus serotypes, human behaviour, and housing characteristics (Rigau-Pérez et al., 1998).

The seasonal nature of transmission may reflect the influence of climate on the transmission cycle (Johansson et al., 2009b). Dengue incidence is usually associated with warmer, more humid weather. Rainfall may influence dengue incidence through the filling of containers out in the open (e.g. old tyres) which create potential breeding sites for the mosquito. More importantly, the breeding of mosquitoes depends on temperature, humidity, the mosquitoes' life expectancy, life-long fecundity, biting activity and virus incubation (Favier et al., 2005). Given favourable climatic conditions for development of

the dengue-carrying mosquito, the urban environment plays a major role in determining transmission rates.

For a dengue epidemic to occur, a large number of mosquitoes are required along with many people with no immunity to one of the four dengue serotypes (DENV-1, DENV-2, DENV-3, DENV-4) and an opportunity for the two to interact. Although a densely populated urban area may be experiencing favourable climate conditions for the transmission of one or more dengue viruses in a given season, the population may have recently been exposed to one of the serotypes and already be immune, preventing the spread of an epidemic. On the other hand, if different serotypes are circulating simultaneously, more severe cases of dengue could occur in people previously infected with another serotype. The minimum number of susceptible persons in a population for maintenance of disease transmission depends on many factors, particularly population immunity and the number of new susceptible persons who enter the population during the period of dengue transmission (Kuno, 1995). Exact estimation of the critical population size that permits endemic transmission is difficult because of the possibility of reintroduction, demographic change, and vector control activity during the study period. The many potential drivers of dengue, both extrinsic, such as climate, and intrinsic, such as population immunity are often difficult to disentangle. This presents a challenge for modelling dengue risk in space and time.

2.3.2 Previous climate-dengue studies

Although many other factors play a crucial role in dengue transmission (e.g. immune status, socio-economic status, living conditions, etc), several studies have focused on the association between climate and dengue incidence for specific countries or cities. Statistical tools used to measure such associations range from graphic assessments and correlation coefficients to linear regression, time series analysis and Bayesian hierarchical models. Some authors have found that time-lagged climate variables of up to two or three months have a statistically significant association with dengue. For example, Schreiber (2001) showed using regression that dengue incidence in San Juan, Puerto Rico (1988-1993) was significantly influenced by climate over at least a 2 month period, using data indicative of the local water budget. Wu et al. (2007) found, using autoregressive integrated moving average (ARIMA) models, that warmer and less humid weather in

Kaohsiung City, Taiwan (1988-2003) were associated with triggering dengue epidemics and the most dominant effect of weather parameters on dengue incidence was at a lag of 2 months. Wu et al. (2009) conducted a spatial analysis to explore relationships between cumulative incidence of dengue fever, climatic and non-climatic factors in Taiwan. Numbers of months with average temperature higher than 18°C per year and degree of urbanization were found to be associated with increasing risk of dengue fever incidence at the municipality level.

In many tropical countries, a positive association between rainfall and dengue incidence has been documented but a statistically significant relationship was not found for other regions (Kuno, 1995). Inconsistent associations between precipitation, temperature and dengue incidence have been reported in the literature. For example, monthly rainfall and temperature was found to be positively associated with dengue incidence in parts of Thailand (Promprou et al., 2005), while temperature has been shown to have a negative effect on dengue incidence in Thailand during the rainy season 1997–2001 (Nakhapakorn and Tripathi, 2005). Dengue has been found to be negatively associated with anomalously high rainfall in Thailand (Thammapalo et al., 2005) and Barbados (Depradine and Lovell, 2004) which may be related to the fact that the *Aedes* mosquito larvae are washed away from containers during heavy downpours (Kelly-Hope and Thomson, 2008). This has potentially important public health implications for disease incidence prediction as above average rainfall does not necessarily imply heightened risk of dengue. However, it is possible that the risk of dengue will be greater following heavy rainfall when mosquitoes can re-establish breeding sites, causing the larval index of dengue *Aedes* mosquitoes to increase (Strickman and Kittayapong, 2002). For example, in Delhi, India an unexpected dengue epidemic in 2003 coincided with the post monsoon period, which was one of the wettest monsoons in 25 years (Chakravarti and Kumaria, 2005).

These contradictory results might be due to the omission of confounding factors in the model and/or failure to account for the annual cycle. Climate and dengue incidence both exhibit strong seasonality. If this is not accounted for in the model, inference of the true predictive relationship between climate and dengue can be misleading. Johansson et al. (2009b) used hierarchical Poisson regression models with monthly reported dengue cases in Puerto Rico (July 1986 - December 2006) regressed on monthly average precipitation and temperature, lagged up to 2 months, with a population offset and a natural cubic spline function of time to adjust for seasonal confounding. The spline likely contains

variation attributable not only to weather. Accounting for this variation in the analysis allows the interpretation of the effects of inter-annual variation in climate on dengue incidence. Although the spline allows a more reliable interpretation of climate-dengue relationships, it is of little use for predictive purposes. Alternative methods to account for seasonal variation in such models include using oscillatory *sine* and *cosine* functions (e.g. Tipayamongkholgul et al., 2009; Fuller et al., 2009) or simply including calendar month as a categorical variable in the model (e.g. Lowe et al., 2010).

Many studies have included some form of ENSO index in statistical models to predict dengue in various locations (e.g. Hales et al., 1996, 1999; Gagnon et al., 2001; Cazelles et al., 2005; Arcari et al., 2007; Brunkard et al., 2008; Hurtado-Diaz et al., 2007; Johansson et al., 2009a; Tipayamongkholgul et al., 2009; Fuller et al., 2009). Hales et al. (1996) found that between 1970 and 1995, the annual number of dengue epidemics that occurred across the island nations of the South Pacific was positively correlated with the SOI. High positive values of the SOI denote La Niña conditions, which are associated with much warmer and wetter conditions than average in the South Pacific. In a subsequent study, the number of dengue cases reported in each of the 14 island nations of the South Pacific was examined (Hales et al., 1999). In general, more cases were reported in warm, wet years than dry cool years, but the authors point out that regional social and environmental factors were not included in the analysis. Arcari et al. (2007) found that a combination of rainfall, temperature, humidity and the SOI can explain 12.9-24.5% of the variance in dengue cases observed in each of eight selected provinces in Indonesia (1992-2001). However, a Gaussian multiple regression model was used to model counts of dengue cases and non-climatic confounding factors were not included in the analysis. Therefore, inference from this model may not be representative of the true relationship between dengue and climate in Indonesia. Brunkard et al. (2008) found, using an autoregressive model, that ENSO, precipitation and temperature played a small yet significant role in dengue transmission in Matamoros, Mexico (1995–2005). Their findings with respect to temperature, precipitation and SST were in general in agreement with the findings of Hurtado-Diaz et al. (2007) with increases in all three climate covariates associated with increased dengue incidence. The model was refitted to data from 1995-2004 and an out-of-sample prediction was obtained for 2005. The authors noted that despite the strong influence of the autoregressive components (dengue cases 1 and 2 weeks lag) in the model, adding the three climate variables resulted in significant improvement in

model fit. However, the predictive power of this model does not extend beyond 2 weeks, offering very little time to implement preventative measures. Johansson et al. (2009a) carried out statistical time-series analyses to examine the dynamic relationship between climate variables and the incidence of dengue in Thailand, Mexico, and Puerto Rico. Geographical variation in the observed impact of ENSO on dengue incidence was found. In Mexico, no association was found, while in Thailand, the association was not statistically significant. In contrast, in Puerto Rico, a statistically significant association was found between 1995 and 2002. However the authors viewed this finding with caution and noted that the role of ENSO may be obscured by local climate heterogeneity, insufficient data, randomly coincident outbreaks, and other, potentially stronger, intrinsic factors regulating transmission dynamics.

Statistical associations between climate and dengue has been noted for several Brazilian cities. For example, Luz et al. (2008) used ARIMA models for dengue incidence in the city of Rio de Janeiro from 1997 to 2004. They found that a 1-step ahead approach for predicting dengue incidence provides significantly more accurate predictions than the 12-steps ahead approach. Temperature and rainfall were found to be positively associated with dengue incidence. Incorporating simultaneous temperature and 1 month lag number of rainy days into the model improved the predictive power of the model, although the improvement was not statistically significant. Câmara et al. (2009) found that temperatures in Rio de Janeiro in the first quarter of the year (1986–2003) were significantly higher in the years in which dengue epidemics started in the city. There was no significant relationship with total rainfall for the same quarter of the year, but epidemics were more frequent in the years in which the volume of rain during the summer was lower than normal. Lowe et al. (2010) presented the first modelling framework for dengue in Brazil at the national and regional level. A spatio-temporal Bayesian hierarchical model was formulated, using climatic and non-climatic covariates and random effects, which allowed probabilistic forecasts to be issued (see subsequent chapters for more details). Such models are useful for developing epidemic early warning systems, as probabilistic forecast uncertainty can be easily quantified. This type of large-scale study allows the identification of potential candidate regions and microregions for the development of a climate based early warning system. The modelling framework is developed, extended and evaluated in this thesis.

The inclusion of covariates based on climate (e.g. temperature, precipitation, Pacific

SST) and their lagged effects appear to be potentially important components of a climate informed dengue prediction model. However, there are inconsistencies in the sign and magnitude of reported statistical associations between climate and dengue. Some models include multiple explanatory variables with multiple possible time lags (e.g. Tipayamongkholgul et al., 2009), which can lead to overfitting (Lafferty, 2009). Few studies have included other non-climatic factors that can affect dengue transmission such as socioeconomic status, levels of urbanization or serotype prevalence (e.g. Hales et al., 1999; Chakravarti and Kumaria, 2005; Câmara et al., 2009) or account for seasonality (e.g. Nakhapakorn and Tripathi, 2005; Promprou et al., 2005). When autoregressive dengue terms are included in the model formulation, the additional gain in variance explained from climate variables is not always quantified (e.g. Tipayamongkholgul et al., 2009; Brunkard et al., 2008). In addition, models for count data are not always employed for modelling dengue cases (e.g. Arcari et al., 2007) and most of the studies mentioned here have not tested models on out-of-sample data (e.g. Schreiber, 2001). There is a need for further research into the potential of using climate in dengue early warning systems for different populations with various climatic/ecological regions, using adequate and well-specified statistical models that are able to capture observed and unobserved confounding factors. This will help to correctly attribute the contribution of climate to variations in dengue.

2.3.3 Dengue in Brazil

Brazil is the largest and most populated country in Latin America, covering more than 8.5 million km^2 with an estimated population of 191.4 million inhabitants in 2009[6]. High population density areas and cities (up to 12,901 inhabitants/km^2) are located mainly on the Atlantic Coast (Siqueira et al., 2005). Large areas of Brazil have highly favourable climate for the proliferation of *Aedes aegypti* and dozens of metropoles with high human population densities living in substandard conditions with deficient sanitation services (Teixeira et al., 2009). Brazil currently accounts for the majority of dengue cases reported in the Americas (Braga et al., 2009). The South East region of Brazil and, in particular, the city of Rio de Janeiro, have been most affected by dengue (Nogueira et al., 2007a; Câmara et al., 2007).

[6] http://www.sidra.ibge.gov.br/ [accessed 15 May 2010]

After the discontinuation of the PAHO *Aedes aegypti* eradication programme in the 1970s, the subsequent re-infestation of *Aedes aegypti* into urban areas of Brazil and the introduction of DENV–1 serotype to Rio de Janeiro in 1986 led to resurgence of dengue fever outbreaks (Schatzmayr et al., 1986). This serotype dispersed throughout Brazil, reaching the North East and North regions. In 1990, DENV–2 was detected, also in Rio de Janeiro (Nogueira et al., 1990), and the first case of dengue hemorrhagic fever was reported (Teixeira et al., 2005). The appearance of DENV–3 is believed to have contributed to a severe epidemic in Brazil in 2002 (Nogueira et al., 2005). DENV-3 predominated until 2005 but later DENV-1 and DENV-2 returned to different parts of the country. In recent years an increase in the number of dengue fever cases amongst children has been observed including severe and fatal cases (Nogueira et al., 2007b). During the 2008 epidemic, there were more than 155,000 cases of dengue in the state of Rio de Janeiro (incidence of 2,544 cases per 100,000 inhabitants), more than 9,000 hospitalisations and 110 deaths, of which nearly half were children (Teixeira et al., 2009). The circulation of three dengue serotypes is likely responsible for the increased occurrence of more severe forms of the disease. In August 2010, the Brazilian newspaper O Globo[7] reported that dengue cases caused by serotype DENV–4 were confirmed in the capital of Roraima in the North of Brazil on the border with Venezuela. The Brazilian Ministry of Health have warned that the propagation of this new serotype through Brazil could cause a huge epidemic, as the majority of the Brazilian population have never been in contact with this virus.

2.3.4 Surveillance and control

Brazil has an excellent laboratory-based surveillance system for dengue and dengue hemorrhagic fever, with a network of laboratories to conduct serological diagnosis and virological surveillance (Gubler, 2002a). However, this system does not perceive changes in incidence early enough for adequate response (Luz et al., 2008). For example, in December 2001, DENV–3 was detected in Rio de Janeiro (Nogueira et al., 2005), but despite warnings, control measures were not implemented until the 2002 epidemic was near its peak transmission. With appropriate coordination and data sharing, this system could be incorporated in a prediction model including socio-economic and environmental risk

[7] http://oglobo.globo.com/vivermelhor/mat/2010/08/12/virus-da-dengue-4-esta-no-brasil-por-populacao-ser-vulneravel-ao-sorotipo-epidemia-causaria-grande-numero-de-mortes-917378354.asp [accessed 15 August 2010]

factors to provide early warnings for epidemic activity. The National Program for the Control of Dengue (PNCD) of Brazil, launched in July 2002 by the Brazilian Ministry of Health, use pre-defined epidemic thresholds to evaluate dengue risk in different areas of Brazil for a given time period. For example, a high dengue incidence rate is defined as 300 cases per 100,000 inhabitants (see Chapter 3, Section 3.2.2). While observing if early cases exceed this threshold may help to detect the early phase of an epidemic, if used in isolation, this approach offers very short lead times (e.g. 1-3 weeks) for preventative control planning and implementation. Therefore the predictive lead time that could be gained by using climate information in a dengue early warning system has potential benefit.

2.3.5 Climate and geography of Brazil

Due to variations in altitude, pressure, air circulation, proximity to the Atlantic Ocean and vegetation type, Brazil can be divided into distinct climatic and geographic zones (see Fig 2.3). The following information was obtained from the Brazilian Institute for Geography and Statistics (IBGE)[8] and the Brazilian government[9]. Brazil has three types of climate (see Fig 2.3a):

- Equatorial: covers primarily the region of the Amazon Rainforest. In equatorial areas it rains almost every day and the average temperature ranges from 25-27°C.
- Tropical (divided into Central Brazil, North East and the Eastern Equatorial Zone): found in the North East, South East and Midwest. Characterized by average temperature above 20°C (in summer temperatures can exceed 25°C) and high rainfall. In winter there may be dry periods.
- Temperate: found in the southern and colder part of the country, south of the Tropic of Capricorn. Rainfall is distributed regularly during the year and the seasons are well defined. Annual mean temperatures are around 18°C. In austral winter, temperatures sometimes fall below zero with possible snowfall.

A biome is a set of vegetation types that covers large contiguous areas on a regional scale, with similar flora and fauna, climate and geography. In Brazil, the defined biomes

[8] http://mapas.ibge.gov.br/, [accessed 20 August 2010]
[9] http://www.brasil.gov.br/sobre/geography/, [accessed 20 August 2010]

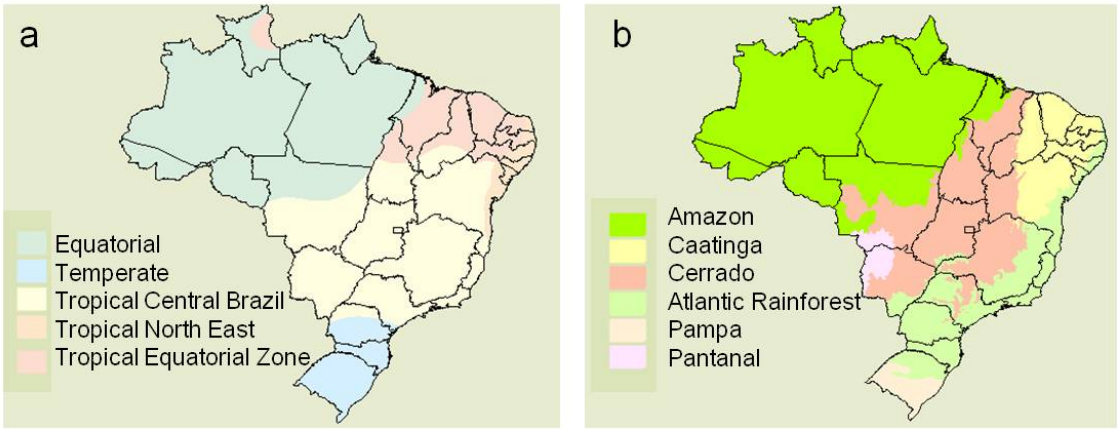


Figure 2.3: The climate and biome zones defined by IBGE *Source:* IBGE, 2010.

are Amazon, Atlantic Rainforest, Caatinga, Pampa and Pantanal (see Fig 2.3b). The Brazilian Amazon is the largest biome in Brazil, occupying nearly half the country and has the largest reserve of biodiversity in the world. It is dominated by hot and humid climate, rivers with permanent heavy flow and vegetation usually consists of tall trees. The Caatinga (indigenous name meaning 'clear and open forest') is an exclusively Brazilian biome that occupies just over 10% of the country. The dry climate, the light and warmth characteristic of tropical Caatinga areas result in a thorny and deciduous (when the leaves fall at a given time) savanna vegetation. The Cerrado biome covers 22% of Brazilian territory and it is dominated by formations of savanna and hot sub-humid tropical climate, a dry season and a rainy season, with an average annual temperature between 22-27°C. Cerrado comprises mountain ranges, valleys, plateaus and plains and its main vegetation consists of tall trees related to a hot and humid climate. The Atlantic Rainforest is an environmental complex that includes mountain ranges, valleys, plateaus and level lands throughout the east Atlantic continental range of Brazil. Its main type of vegetation is tropical rain forest, usually consisting of tall trees and related to a hot and humid climate. The Pampa biome is marked by a rainy climate without a dry season, regular polar fronts and freezing temperatures in winter. The predominant vegetation consists of herbs and shrubs. Finally, the Pantanal is a biome characterized by floods of long duration (due to the low permeability of the soil) and its predominant vegetation is the savannah.

2.3.6 Climate variations in Brazil

To understand how variations in climate might effect dengue incidence in Brazil it is important to first be aware of atmospheric teleconnections that affect patterns of weather variability. Teleconnection is a term used to describe the tendency for atmospheric circulation patterns to be related, either directly or indirectly, over large and spatially non-contiguous areas (Bridgman and Oliver, 2006). Rainfall and temperature anomalies associated with occurrence of El Niño and La Niña events are the major source of inter-annual variability over much of South America (Garreaud et al., 2009). El Niño is associated with high surface pressure over the western tropical Pacific, low surface pressure over the southeastern tropical Pacific and an eastward displaced Walker circulation with an anomalous rising motion over central and eastern equatorial Pacific (Philander, 1990). This coincides with heavy rainfall, unusually warm surface waters and relaxed trade winds in the central and eastern tropical Pacific. During La Niña, surface pressure is high over the eastern but low over the western tropical Pacific. Trade winds are intensified and SST and rainfall are low in the central and eastern tropical Pacific. The SST anomalies associated with El Niño and La Niña episodes produce anomalous heat and water vapor fluxes from the tropical Pacific Ocean to the atmosphere (Grimm and Tedeschi, 2009). The associated convection anomalies cause upper-level divergence anomalies that perturb the global circulation. In Brazil, El Niño events tend to coincide with reduced Amazon rainfall, especially in the northern and central regions, while opposite anomalies often occur during La Niña events (Liebmann and Marengo, 2001). During El Niño, anomalous ascending motion over the eastern equatorial Pacific is thought to produce anomalous subsidence east of the Andes, resulting in below-average precipitation in the northern Amazon (e.g., Marengo and Hastenrath, 1993). El Niño events are also thought to affect the circulation of the Atlantic atmosphere-hydrosphere system via an 'atmospheric bridge' from the equatorial eastern Pacific to the northern tropical Atlantic which in turn affects rainfall variability in North East Brazil (Hastenrath, 2006). During an El Niño event, droughts in the North East can occur when the Intertropical Convergence Zone (ITCZ), the main rain bearing system for the North East, remains anomalously far North. During El Niño, negative precipitation anomalies in North, Central East and North East Brazil are accompanied by positive anomalies in South (Grimm, 2003) and South East (Coelho et al., 2002) Brazil, possibly influenced by the intensification of the subtropical jet stream (Philander, 1990). Such precipitation anomalies are

favored by the perturbation in the Walker and Hadley circulation over the east Pacific and South America, and by a Rossby wave train over southern South America that originates in the eastern Pacific and influences the subtropics (Grimm, 2003). Due to the relatively high predictability of the ENSO phenomenon, prior knowledge of the expected state of the equatorial Pacific Ocean provides a significant source of predictability of seasonal climate variability over much of the tropics (Mason and Goddard, 2001). Due to the time lags of the order of several months involved between the effect of SSTs on regional climate variability and the subsequent influence on disease risk, predictive lead time can be gained by including ENSO information in a disease early warning system.

2.4 Summary

Dengue has been identified as an important climate-sensitive disease. Due to the large dengue burden in Brazil (Luz et al., 2009) and increased severity of the disease over the last two decades (Teixeira et al., 2009), Brazil could benefit from a dengue early warning system. This chapter provides a review of current state of knowledge and serves as a platform to base further development of a predictive model for disease risk. Several important aspects should be considered in the analysis of predictive models for early warning of climate-sensitive disease risk. These include:

- Use of climate and non-climate information to correctly attribute variation in disease risk to climatic factors.
- Development of spatio-temporal models to predict geographically specific epidemic early warnings.
- Thorough evaluation of the skill of disease risk warnings.

Further research is required to incorporate more sophisticated modelling techniques into climate related dengue studies. Such considerations include:

- The inclusion of random effects to account for unobserved confounding factors.
- The development of probabilistic forecasts to allow prediction uncertainty to be quantified.

In the chapters that follow, a statistical framework is developed to model spatio-temporal variation in dengue risk in Brazil from January 2001 - December 2009 using both climatic and non-climatic variables. After extensive exploratory analysis and model selection, the selected model for Brazil is refined in the context of the South East region of Brazil. The impact of unobserved confounding factors are accounted for using a combination of spatially structured and unstructured random effects in a Bayesian hierarchical model, estimated using Markov Chain Monte Carlo (MCMC). This methodology allows posterior predictive distributions for dengue risk to be derived. Probabilistic predictions for dengue epidemics in South East Brazil are then evaluated against out-of-sample data, for the peak dengue season in 2008 and 2009, to assess the potential of incorporating the model into a dengue epidemic early warning system. The model developed in this thesis is compared to a simple model representative of current practice in dengue surveillance in Brazil and the potential benefit of combining these two forecasting systems is considered.

Chapter 3

Exploratory data analysis

3.1 Introduction

The aim of this chapter is to introduce the datasets that will be used in this study to develop a spatio-temporal statistical modelling framework to predict dengue risk in Brazil. Data were obtained through both access to standard website sources and from research visits to Brazilian institutes. This facilitated the collection of climate, cartographic, demographic and disease datasets and expert knowledge on current public health monitoring and surveillance for dengue in Brazil. Institutes included:

- Center for Weather Forecasting and Climate Studies (CPTEC)
- National Institute for Space Research (INPE)
- Oswaldo Cruz Foundation (FIOCRUZ)
- Brazilian Institute for Geography and Statistics (IBGE)

3.2 Dengue data

Dengue fever data (counts of notified cases per calendar month) from January 2001 - December 2009 (108 months) were obtained at municipality level (5564 municipalities) from DATASUS[1] (Unified Health System Database), established by the Brazilian Min-

[1] http://dtr2004.saude.gov.br/sinanweb/novo/ [accessed 15 May 2010]

istry of Health. The dataset includes all notified dengue cases from hospitals and clinic doctors from both the private and public health system. An example of the form to be completed by clinicians to notify SINAN (Information System for Notifiable Diseases) of a new dengue case is presented in Figure 3.1. This includes information on basic demographic data, dates of symptom onset and sample collection and case classification (e.g. classic dengue fever, dengue hemorrhagic fever, dengue shock syndrome). Individual data are locally entered into the electronic information system and subsequently transmitted to state and national levels (Siqueira et al., 2005). Each entry represents one case event. Cases are laboratory confirmed where possible, or otherwise based on syndromic definition[2]. During the January 2001 - December 2009 period, 3.25 million cases of dengue were reported across Brazil.

A network of laboratories, capable of diagnosing dengue infections, has been implemented in all Brazilian states. The network is responsible for confirmation of cases to support epidemiological surveillance (Nogueira et al., 2007b). However, this network is not accessible to all municipalities. To address this issue, dengue counts were aggregated to the microregion level (558 microregions), where a microregion typically consists of one large city and several smaller municipalities. This alleviates problems of misreporting due to variation in availability of health services/epidemiological facilities at the municipality level. In Figure 3.2a, time series of dengue counts for the 2001-2009 period grouped into the five main regions of Brazil (see Fig. 3.2b) are presented. Table 3.1 summarises the total dengue cases reported in each region and Figure 3.2c shows the total dengue cases in each microregion over the period January 2001-December 2009. Dengue is most prevalent in the South East. Two major epidemics occurred in the late austral summer of 2002 and 2008, while considerably fewer dengue cases were reported in 2004 and 2005 (Fig. 3.2a). Figure 3.2c shows that there are fewer dengue cases in South Brazil and the North West Amazon. Given this dataset, the mean dengue cases per month in any given microregion was 54 with a variance of 432,213. Note that the variance is in excess of the mean which has implications for the consideration of Poisson models in subsequent chapters.

[2]http://portal.saude.gov.br/portal/arquivos/pdf/notasriehitricadenguemaro2009.pdf [accessed 15 May 2010]

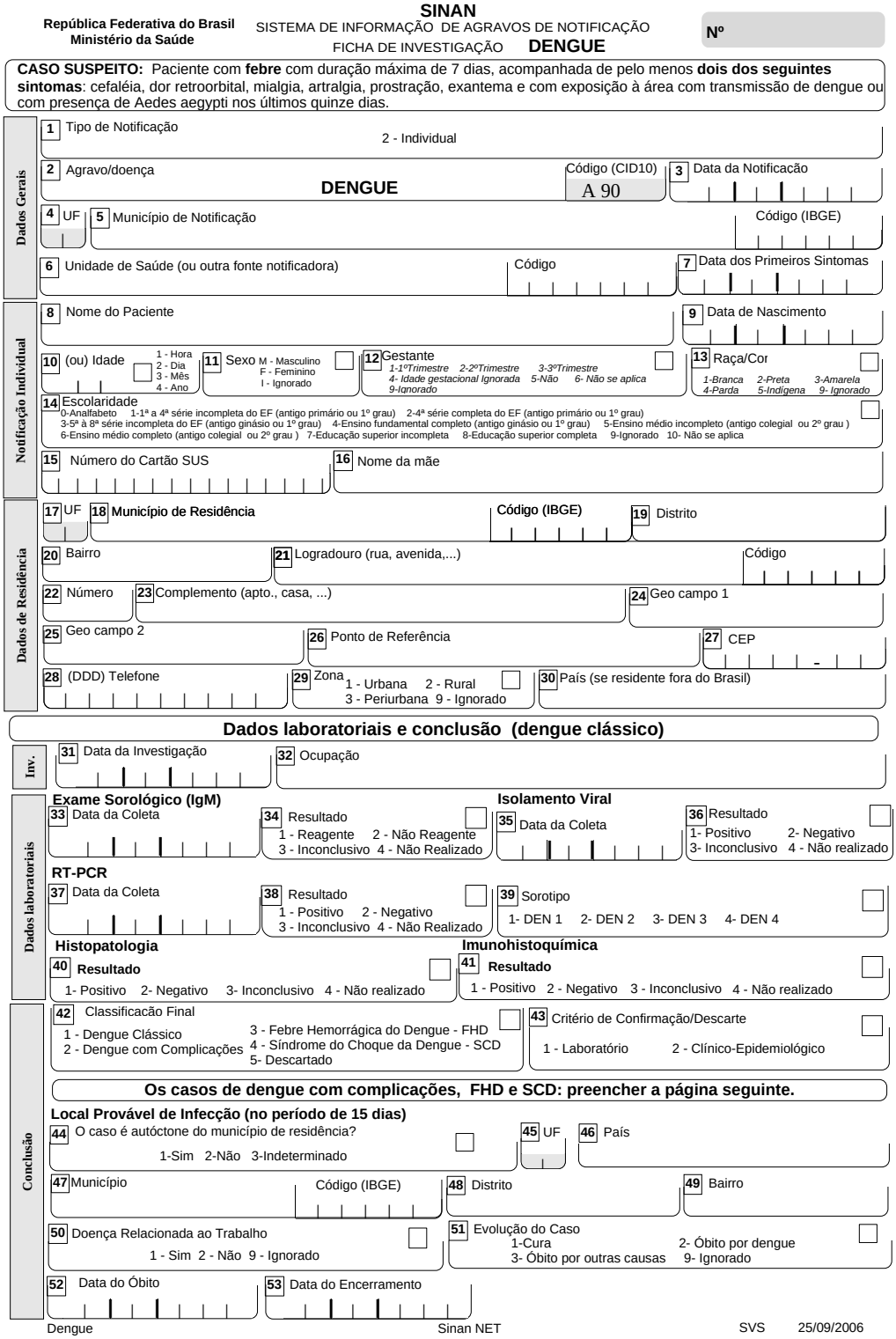


Figure 3.1: Form completed by clinicians when a patient is suspected to have dengue.

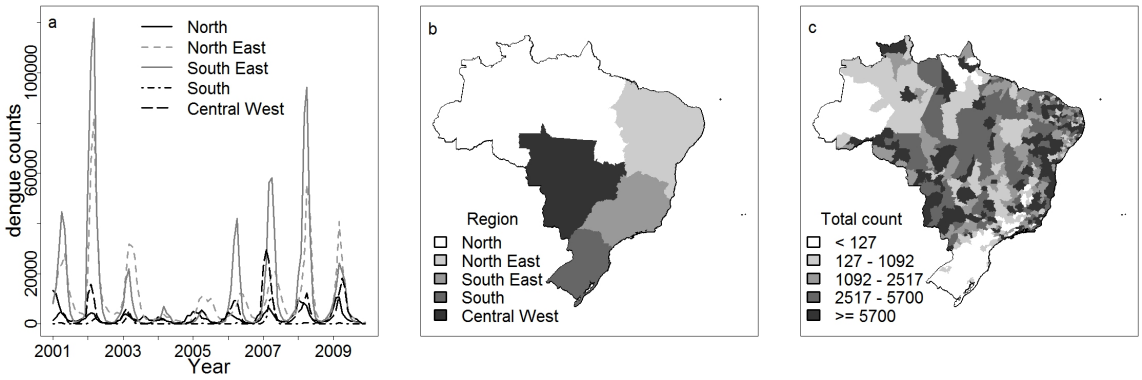


Figure 3.2: (a) Monthly dengue counts for main regions of Brazil from January 2001 - December 2009. Map to show (b) main regions of Brazil, (c) total dengue cases in each microregion for the given time period.

Table 3.1: Notified dengue cases and dengue incidence rate for period January 2001 - December 2009 for main regions of Brazil.

Region	Total notified dengue cases	Overall dengue incidence rate[a]
North	293,120	225
North East	1,167,116	255
South East	1,345,773	192
South	55,422	23
Central West	389,159	335

[a]dengue incidence rate defined as the number of dengue cases per 100,000 inhabitants.

3.2.1 Limitations of dengue data

The quality of the dengue dataset depends on the technical and operational system of epidemiological surveillance in each geographic area to detect, report, investigate and conduct specific laboratory tests to confirm the diagnosis of dengue cases. Underreporting may be due to non-declared or self-diagnosed cases or attributed to difficulties in identifying the clinical forms of mild and moderate forms of the virus, which constitute the majority of cases of dengue. During epidemics, the health services are overwhelmed by patients with suspected dengue, therefore not all cases are notified or laboratory confirmed. In some instances overestimation occurs. Dengue symptoms are similar to those of other diseases such as leptospirosis (Flannery et al., 2001) and mis-diagnosis is common, particularly during dengue epidemics when case confirmation is based on

symptoms alone. Because of the above, the dataset will contain errors concerning the exact magnitude and timing of an epidemic. However, the occurrence of an epidemic can be detected using this dataset, as evidenced by published literature, media coverage and archives provided by ProMED mail, a Program for Monitoring Emerging Diseases[3].

A further disadvantage is that the data provided is not broken down by virus type. Serological information could be useful to indicate the periodicity of circulating serotypes (DENV-1, DENV-2, DENV-3, DENV-4) which influence population immunity and hence the occurrence of epidemics. Further, as temperature and precipitation influence the abundance and transmission potential of *Aedes aegypti*, it would be advantageous to include entomological data in the analysis. However, this information was unobtainable.

In addition to the limited epidemiological dataset, other issues arise in relation to the grouping of population into large and spatially extensive microregions (microregion area ranges from 17 km^2 to 332,000 km^2) and their association with the measured explanatory variables, which are available at finer or coarser spatial scales than the microregion level (e.g. climate, levels of urbanisation, altitude, see Table 3.2). The aggregation of cases to the microregion level causes smaller spatial scale information to be lost, such as clustering of dengue cases at the sub-microregion level. Therefore the challenge is to construct a predictive model for dengue epidemics using a limited dataset at a coarse spatial resolution. The relationships between dengue and available explanatory variables, outlined in Table 3.2, are explored in the following sections.

[3] http://www.promedmail.org [accessed 15 May 2010]

Table 3.2: Source and original resolution of datasets.

Data	Spatial resolution	Temporal resolution	Source
Dengue cases	Municipality	Monthly count	DATASUS http://dtr2004.saude.gov.br/sinanweb/novo/
Area[a]	Municipality		DATASUS http://www2.datasus.gov.br/DATASUS
Altitude	Municipality		DATASUS http://www2.datasus.gov.br/DATASUS
Population	Municipality	Yearly estimate	IBGE http://www.ibge.gov.br/
Percentage of urban population	Microregion		SIDRA http://www.sidra.ibge.gov.br/
Households with at least one bathroom	Microregion		SIDRA http://www.sidra.ibge.gov.br/
Refuse collection	Microregion		SIDRA http://www.sidra.ibge.gov.br/
Water supply provided by a network	Microregion		SIDRA http://www.sidra.ibge.gov.br/
Biome	Biome		IBGE, provided by Christovam Barcellos
Precipitation rate	$2.5^\circ \times 2.5^\circ$ grid	Monthly mean	GPCP data provided by NOAA/OAR/ESRL PSD, http://www.esrl.noaa.gov/psd/
Surface temperature	$2.5^\circ \times 2.5^\circ$ grid	Monthly mean	NCEP/NCAR Reanalysis data provided by NOAA/OAR/ESRL PSD, http://www.esrl.noaa.gov/psd/
Oceanic Niño Index		3-month running mean	NOAA CPC http://www.cpc.noaa.gov/

[a]mean municipality area = 1,500 km^2, mean microregion area = 14,200 km^2, mean biome area = 42,548,632 km^2, approximate grid square area = 77,000 km^2.

3.2.2 Dengue incidence rate

The Brazilian Ministry of Health defines a dengue incidence rate (DIR) as the number of new dengue cases y_s per 100,000 inhabitants for a geographical area in a specified period of time[4].

$$\text{DIR} = \frac{y_s}{p_s} \times 100,000 \tag{3.1}$$

where p_s is the population in a given time period. In order to calculate incidence rates using the dengue count dataset described above, yearly population estimates for Brazilian municipalities from 2001-2009 were obtained from the Brazilian Institute for Geography and Statistics (IBGE)[5]. These estimates are based on the 2000 census and take into account changing demographic components such as births, mortality and migration. Microregion population estimates were found by calculating the sum of the population estimates for the municipalities located in each microregion. Figure 3.3a illustrates a slight upward trend in population for the main regions of Brazil from 2001-2009, particularly in the South East region. Figure 3.3b shows the spatial distribution in 2009, highlighting the location of the main metropolises in Brazil (e.g. Brasilia, Rio de Janeiro, São Paulo, Manaus), with a population greater than 1 million inhabitants. The estimated total population for Brazil in 2009 was 191.4 million inhabitants. In order to calculate population density estimates, data for the area of each municipality were obtained from DATASUS[6]. The area of each microregion was found by calculating the sum of the area of the municipalities located in each microregion. Population density estimates were calculated by diving microregion population by microregion area.

The National Program for the Control of Dengue (PNCD)[7] in Brazil classifies areas of Brazil (e.g. regions, states, microregions, municipalities) according to three dengue risk categories:

- Low incidence - incidence rate less than 100 cases per 100,000 inhabitants
- Medium incidence - incidence rate between 100 and 300 cases per 100,000 inhabitants
- High incidence - incidence rate greater than 300 cases per 100,000 inhabitants.

[4] http://www.ripsa.org.br/fichasIDB/record.php?lang=pt&node=D.2.3 [accessed 15 May 2010]
[5] http://www.ibge.gov.br/ [accessed 15 May 2010]
[6] http://www2.datasus.gov.br/DATASUS, [accessed 15 May 2010]
[7] http://portal.saude.gov.br/portal/arquivos/pdf/pncd_2002.pdf [accessed 15 May 2010]

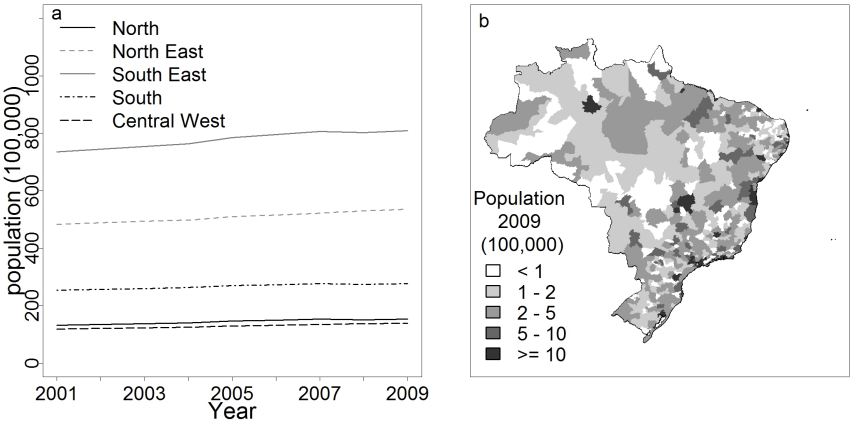


Figure 3.3: (a) Yearly population estimates for main regions of Brazil 2001 - 2009. (b) Total population in 2009 for each microregion.

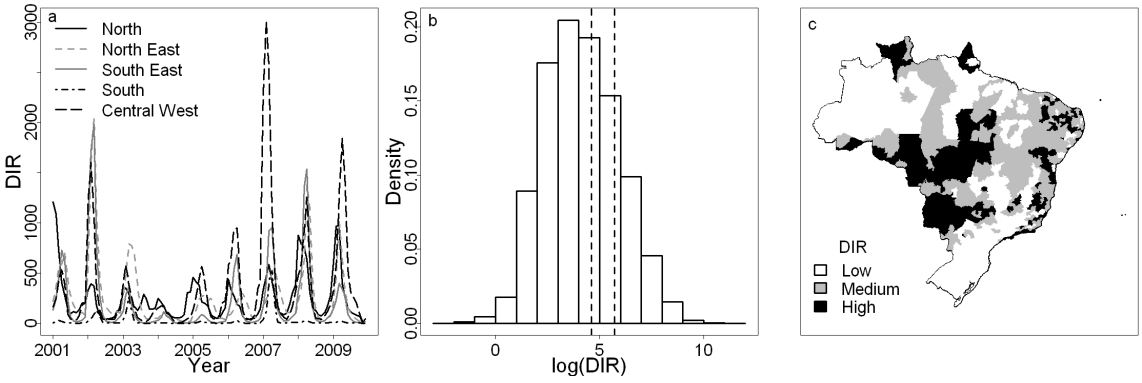


Figure 3.4: (a) DIR for main regions of Brazil January 2001 - December 2009. (b) Histogram of log(DIR). Dashed vertical lines represent risk thresholds of 100 and 300 cases per 100,000 inhabitants. (c) Map of low (less than 100), medium (between 100 and 300) and high (greater than 300) dengue incidence in each microregion over 2001-2009.

In Figure 3.4a, time series of the DIR for the 2001-2009 period, grouped into the five main regions of Brazil, are presented (total region DIR presented in Table 3.1). Figure 3.4b shows a corresponding histogram of DIR (note log scale) and the defined dengue risk thresholds of 100 and 300 cases per 100,000 inhabitants. Figure 3.4c illustrates the spatial distribution of DIR according to the three risk categories; high, medium and low incidence. It is interesting to note the Central West region experienced a dengue epidemic in 2007 and much of this region experienced 'high' dengue incidence for the time period. However, this is not as apparent when examining the raw dengue counts alone (see Fig. 3.2a) as this region has the lowest population of the five main regions.

3.2.3 Standardised morbidity ratios

In the analysis of disease counts within a geographic area, disease maps typically show standardised mortality or morbidity (e.g. incidence) ratios for geographic areas (Elliott and Wartenberg, 2004). The standardised morbidity ratio (SMR) is defined as the ratio of observed dengue cases y_{st} to the expected number of cases e_{st} within a microregion s at time t, where $s = 1, \ldots, S$, with $S = 558$ and $t = 1, \ldots, T$, with $T = 108$ (108 months in the dataset). The expected cases e_{st} in each microregion are calculated as the population $p_{st^*(t)}$ within a microregion s at year $t^*(t)$, where $t^*(t) = 1, \ldots, 9$, multiplied by the overall annual observed risk π, i.e. the total number of cases (over all microregions s and months t) divided by the total population over the time period. Note that as the temporal resolution of the population data is yearly rather than monthly, the population estimate is divided by 12 to calculate monthly expected cases. This allows the comparison of monthly with yearly dengue ratios (see Eqn. 3.2). A ratio greater than 1 in a given time period would suggest an excess risk of dengue fever in the microregion.

$$\begin{aligned} \text{SMR}_{st} &= \frac{y_{st}}{e_{st}} \\ e_{st} &= (p_{st^*(t)}/12)\pi \\ \pi &= \frac{\sum_{s=1}^{S}\sum_{t=1}^{T} y_{st}}{\sum_{s=1}^{S}\sum_{t=1}^{T} p_{st^*(t)}}. \end{aligned} \tag{3.2}$$

Since dengue relative risks are of interest, population effects are eliminated by including the expected number of cases in each microregion as an offset in models used later in this thesis. Therefore, it is essentially the SMR that is modelled for each microregion in a given time period. The SMR, as described above, is based on internal standardization, specific to Brazil. Some authors (Julious et al., 2001) advocate standardization which involves adjustment to a common standard (e.g. incidence rate per 100,000 population). However, as both methods are directly proportional to one another, they give similar results. The overall risk for Brazil was found to be $\pi = 0.002$, which gives equivalent risk thresholds of SMR=0.5 and SMR=1.5 (100 and 300 cases per 100,000 inhabitants respectively, based on PNCD classification). For ease of interpretation all subsequent results will be presented as dengue incidence rates per 100,000 inhabitants (Eqn. 3.1); the standard adopted by the Brazilian Ministry of Health.

3.3 Cartographic and demographic data

National cartographic data such as altitude, area and biome (i.e. climatically and geographically defined zones, see Chapter 2, Section 2.3.5) were obtained from IBGE. Census data at the microregion level such as the percentage of urban population, households with at least one bathroom, refuse collection and water supply provided by a network, were obtained from an aggregated database, SIDRA[8] maintained by IBGE. Poor sanitation conditions such as households without a bathroom, lack of refuse collection services and water supply, encourage mosquito breeding sites. Accordingly, dengue risk is expected to increase as sanitation conditions deteriorate. However, the proportion of the population with improved services increases with levels of urbanisation (see Fig 3.5). Therefore, at the microregion level these variables are not useful for determining dengue risk as they act as a surrogate for the level of urbanisation. While improved services that accompany increased levels of urbanisation reduce dengue risk, the proportion of the population residing in urban areas is expected to increase dengue risk, as urban areas are ideal environments for mosquitoes and many people living in close proximity create a human virus reservoir.

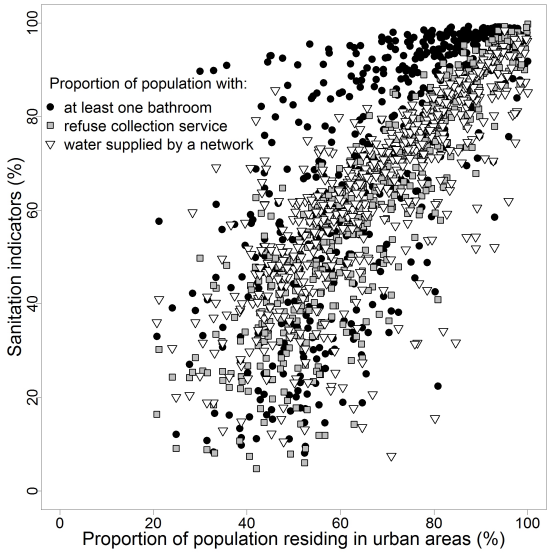


Figure 3.5: Relationship between proportion of population living in urban areas in each microregion and proportion of population with at least one bathroom, refuse collection and water supply provided by a network.

[8] http://www.sidra.ibge.gov.br/ [accessed 15 May 2010]

Each microregion belongs to an administrative main region (1. North, 2. North East, 3. South, 4. South East, 5. Central West) and a biome (1. Amazon Rainforest, 2. Caatinga, 3. Cerrado, 4. Atlantic Rainforest, 5. Pampa, 6. Pantanal). A spatial variable named zone was defined according to the six biomes but with the Atlantic Rainforest biome additionally subdivided into three areas (North East, South East and South) according to different climatic regimes. For example, south of the Tropic of Capricorn (23.5°S) the climate is more temperate and humid, while in the North East portion of the Atlantic Rainforest the climate is relatively warmer and drier. Therefore, eight zones are defined for which climatic, geographical and ecological conditions are approximately homogeneous. An interesting issue is to detect whether there are particular zones where dengue relative risk requires specific modelling. Therefore, zone $k(s)$ where $k(s) = 2, \ldots, 8$ and k is a function of space s, will be considered in models as a categorical factor in subsequent chapters. The Amazon Rainforest zone ($k(s) = 1$) will be set as a reference level. Zone will be interacted with other variables of interest, which may be expected to behave in a way specific to the geographical setting.

The spatial distribution of altitude and urban population in Brazil and the location of the geographical zones are illustrated in Figures 3.6 a, b and c respectively. Figure 3.7 illustrates the dependence of the DIR on these covariates for the given time period (January 2001 - December 2009). DIR has a negative association with altitude (Fig. 3.7a) and a weak positive association with percentage of urban population (Fig. 3.7b). The DIR is lower in zones located in South Brazil (Pampa, South Atlantic Rainforest) and greater in the remaining zones, located north of the Tropic of Capricorn (Fig. 3.7c).

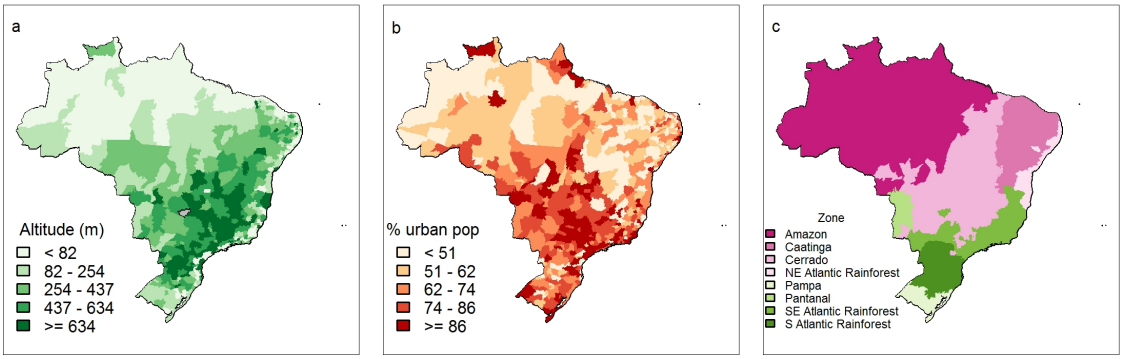


Figure 3.6: Spatial distribution of (a) altitude, (b) urban population, (c) geographic zones in Brazil.

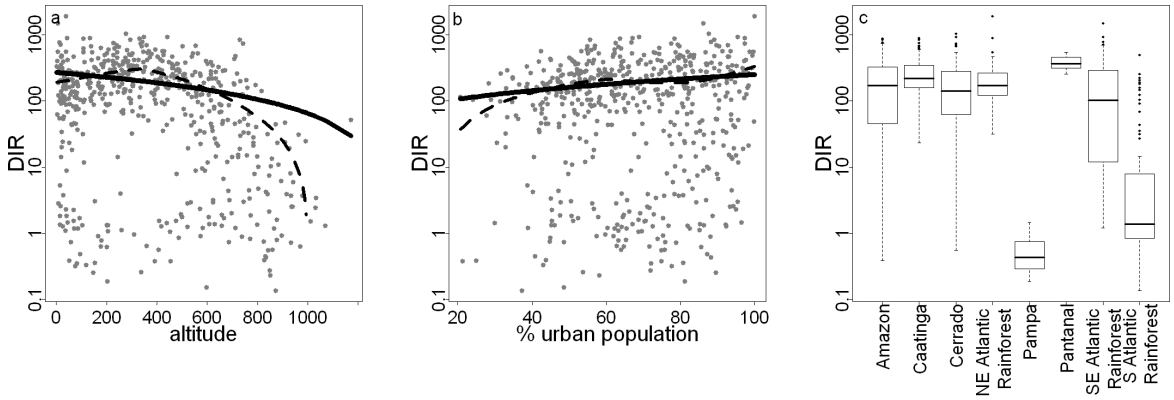


Figure 3.7: Scatter plot between DIR and (a) altitude, (b) % urban population. Solid curve - linear model fit, dashed curve - local polynomial regression fit. (c) Boxplots to show distribution of DIR in each zone. Note logarithmic y axes.

Figure 3.8 shows that DIR has a marked annual cycle which differs between zones. To allow for this, the annual cycle will be included in models in Chapter 4, as a categorical variable for calendar month $t'(t)$, where $t'(t) = 2, \ldots, 12$ and t' is a function of time t. August ($t'(t) = 1$) will be set as a reference level, as DIR for this month is generally low (see Fig. 3.8). DIR peaks between February and April in all zones. The peak occurs in March for Cerrado, Pantanal and Atlantic Rainforest zones (see Fig. 3.8c, d, f, g, and h). The possibility of a spatially varying dengue annual cycle will be allowed for in such models by an interaction term between the categorical variables zone $k(s)$ and calendar month $t'(t)$ (see Chapter 4). As only part of the annual cycle in DIR may be attributable to climatic conditions, the inclusion of this interaction could account for other confounding variables, such as seasonal population movements, leading to zonal differences in the annual cycle.

3.4 Climate data

This section presents the climate data that will be tested as potential explanatory variables for dengue relative risk. Due to the nature and availability of both dengue and climate data for Brazil (see Table 3.2), the analysis will be conducted at a relatively coarse resolution. Therefore, the models formulated in subsequent chapters will be unable to capture sub-microregion variations in dengue which are likely influenced by localised

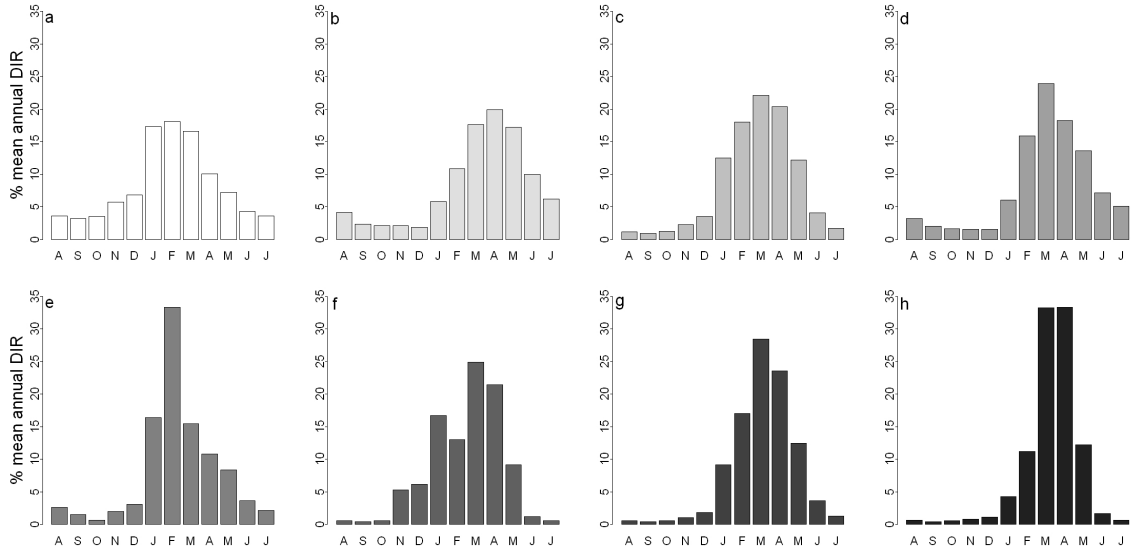


Figure 3.8: Annual cycle of DIR for (a) Amazon Rainforest, (b) Caatinga, (c) Cerrado, (d) North East Atlantic Rainforest, (e) Pampa, (f) Pantanal, (g) South East Atlantic Rainforest and (h) South Atlantic Rainforest, calculated for the period 2001-2009.

meteorological conditions. Rather, the aim of this analysis is to identify any large scale variations in dengue that could be attributed to seasonal variations in temperature and precipitation which are, in part, driven by the El Niño Southern Oscillation.

3.4.1 Precipitation and temperature gridded datasets

Observed gridded ($2.5° \times 2.5°$ latitude-longitude grid) average monthly rate of precipitation data (mm/day) were obtained from the Global Precipitation Climatology Project (GPCP) (Adler et al., 2003). The dataset is a combination of gauge observations with satellite estimates from 1979 to present. The combined dataset was developed and computed by the NASA/Goddard Space Flight Center's Laboratory for Atmospheres as a contribution to the GEWEX Global Precipitation Climatology Project. Reanalysis gridded ($2.5° \times 2.5°$ latitude-longitude grid) monthly mean surface air temperature data (°C) were obtained from the NCEP/NCAR Reanalysis. The NCEP/NCAR Reanalysis project uses a state-of-the-art analysis/forecast system to perform data assimilation using past data from 1948 to the present (Kalnay et al., 1996).

Precipitation and temperature data from both the GPCP combined rain gauge-satellite

dataset and the reanalysis project were extracted for the period 2000-2009 and will be referred to as 'observed' climate variables for the remainder of the text. Due to the availability of dengue data (2001-2009), the climate exploratory data analysis will be conducted for the same time period, as the models developed in subsequent chapters will be conditioned on data for this time period only. Figure 3.9 shows the average precipitation rate and temperature climatology for the December-February (DJF) season (2000-2009). This season was selected as preliminary analyses indicated a lagged correlation between this period and the peak dengue month of March in most zones (see Fig. 3.8). During this season, the highest rainfall occurs in the Amazon Rainforest. An area extending from the South East of Brazil to the North West receives more rainfall than the North East region (Fig 3.9a). During this season the South and South East regions are cooler on average than the rest of Brazil (Fig 3.9b). Figures 3.10 and 3.11 show precipitation and temperature anomalies for the DJF season from 2000-01 to 2008-09 relative to the DJF 2000-01 to 2008-09 average, respectively. Notable precipitation anomalies for the DJF season include the following: below average precipitation across South East Brazil in 2000-01 (Fig 3.10a), above average precipitation along the east coastal area of Brazil 2001-02 (Fig 3.10b), above average precipitation in North East Brazil in 2003-04 (Fig 3.10d), below average precipitation in North and South East Brazil and above average precipitation in North West Brazil in 2005-06 (Fig 3.10f), above average precipitation in South East Brazil in 2006-07 (Fig 3.10g) followed by near average/below average precipitation in this region in 2007-08 (Fig 3.10h) and above average precipitation in North West Brazil, 2008-09 (Fig 3.10i). Temperature anomalies include below average temperatures across most of Brazil in 2000-01 (Fig 3.11a), above average temperatures in South Brazil in 2002-03 (Fig 3.11c), above average temperatures across Brazil in 2006-07 (Fig 3.11g), followed by near average-below average temperatures in 2007-08 (Fig 3.11h).

3.4.2 Comparing gridded datasets to microregion data

Microregion and gridded data were compared by assigning a grid point to each microregion on the basis that the microregion centroid is contained within the grid square, i.e. by calculating the shortest Euclidean distance between microregion centroid and neighbouring grid points (see Fig. 3.12). The response variable dengue counts per month y_{st}, was formulated into a $s \times t$ matrix where $s = 1, \ldots, S$, with $S = 558$ and $t = 1, \ldots, T$, with

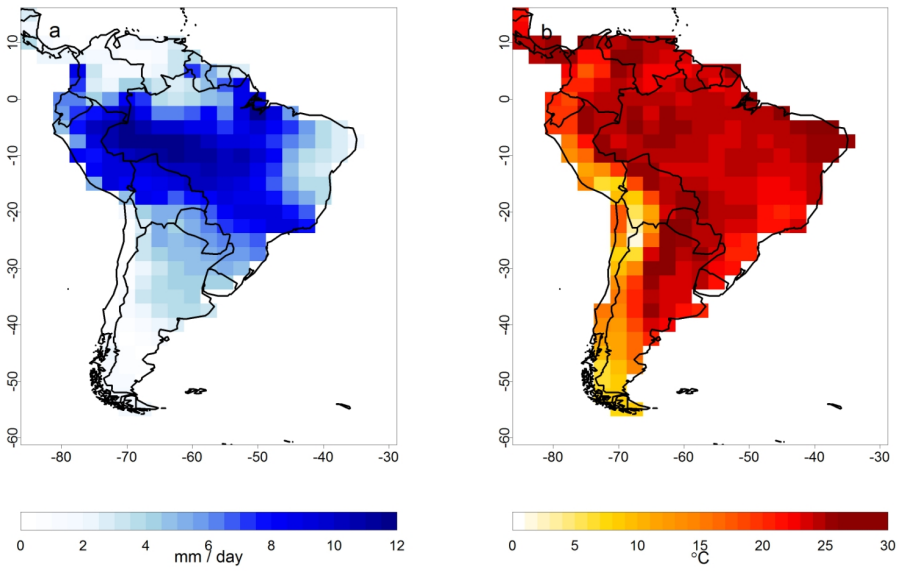


Figure 3.9: (a) Average precipitation rate and (b) temperature climatology in South America for DJF season 2000-2009.

$T = 108$. The corresponding explanatory variables, both climatic x_{jst} and non-climatic w_{jst}, where formulated into a three dimensional array where $j = 1, \ldots, p$ where p is the total number of climatic or non-climatic explanatory variables.

This nearest neighbour interpolation method simply selects the climate variables of the grid square within which the microregion is located (or for large microregions, the gird square closest to the microregion centroid) and does not consider values at other neighbouring grid squares. A limitation to this approach is that a microregion located on the edge of a grid square may experience climatic conditions more similar to adjacent squares. Further, as the microregions are irregularly shaped, the centroid may not be the most representative location of the microregion. A more desirable approach would be to perform statistical downscaling of the gridded climate data to station data available within the microregions. However, no such data were readily available. For the purpose of gaining an understanding of the large-scale relationship between dengue and monthly/seasonal variations in climate at the microregion level, this crude approach was considered acceptable.

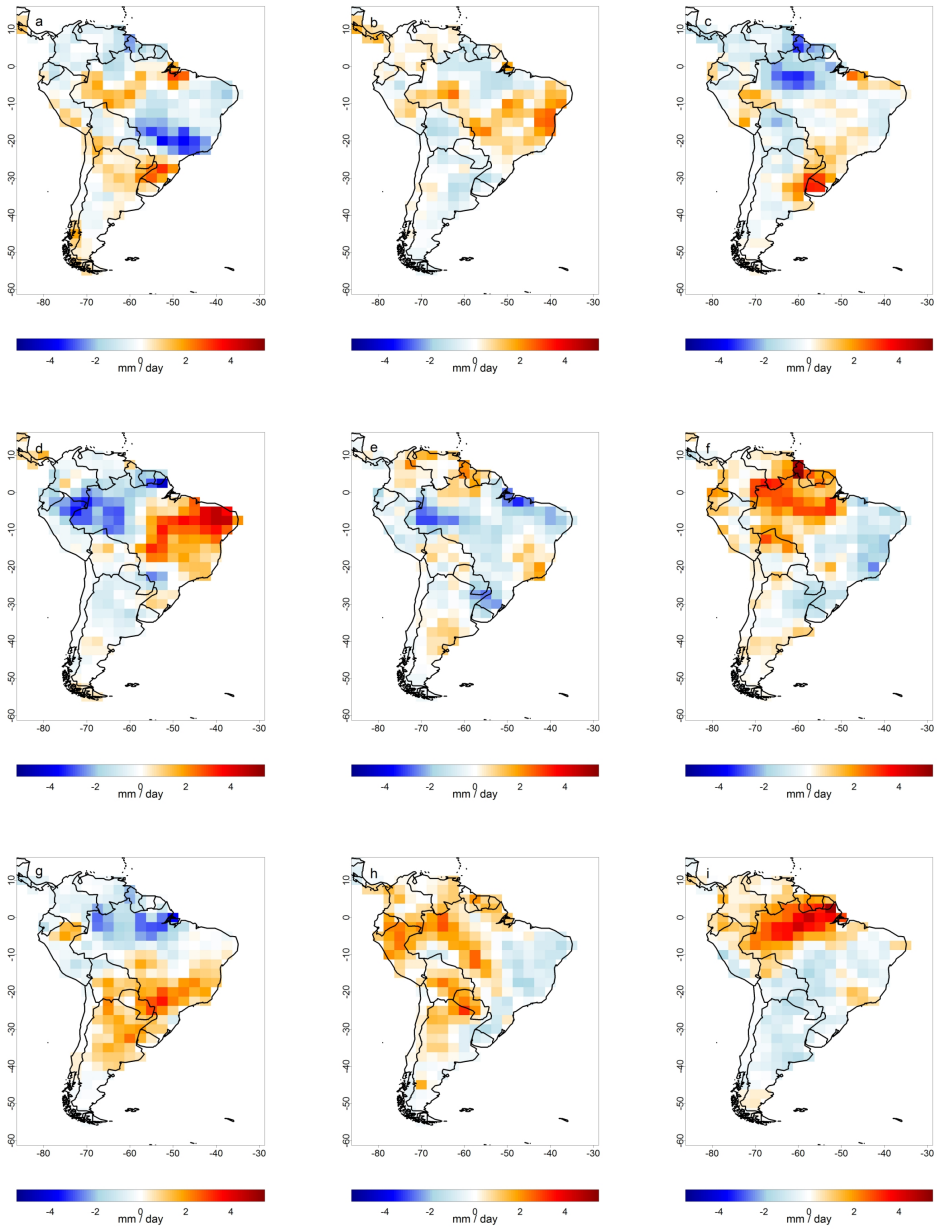


Figure 3.10: Precipitation anomalies, relative to DJF 2000-01 to 2008-09 climatology, in South America for DJF season in (a) 2000-01, (b) 2001-02, (c) 2002-03, (d) 2003-04, (e) 2004-05, (f) 2005-06, (g) 2006-07, (h) 2007-08 and (i) 2008-09.

3.4.3 Dengue and gridded climate

To investigate the relationship between precipitation, temperature and DIR, two zones with contrasting climatic regimes are considered. Firstly, the South East Atlantic Rainforest, which experiences a subtropical climate and has substantial altitude differences

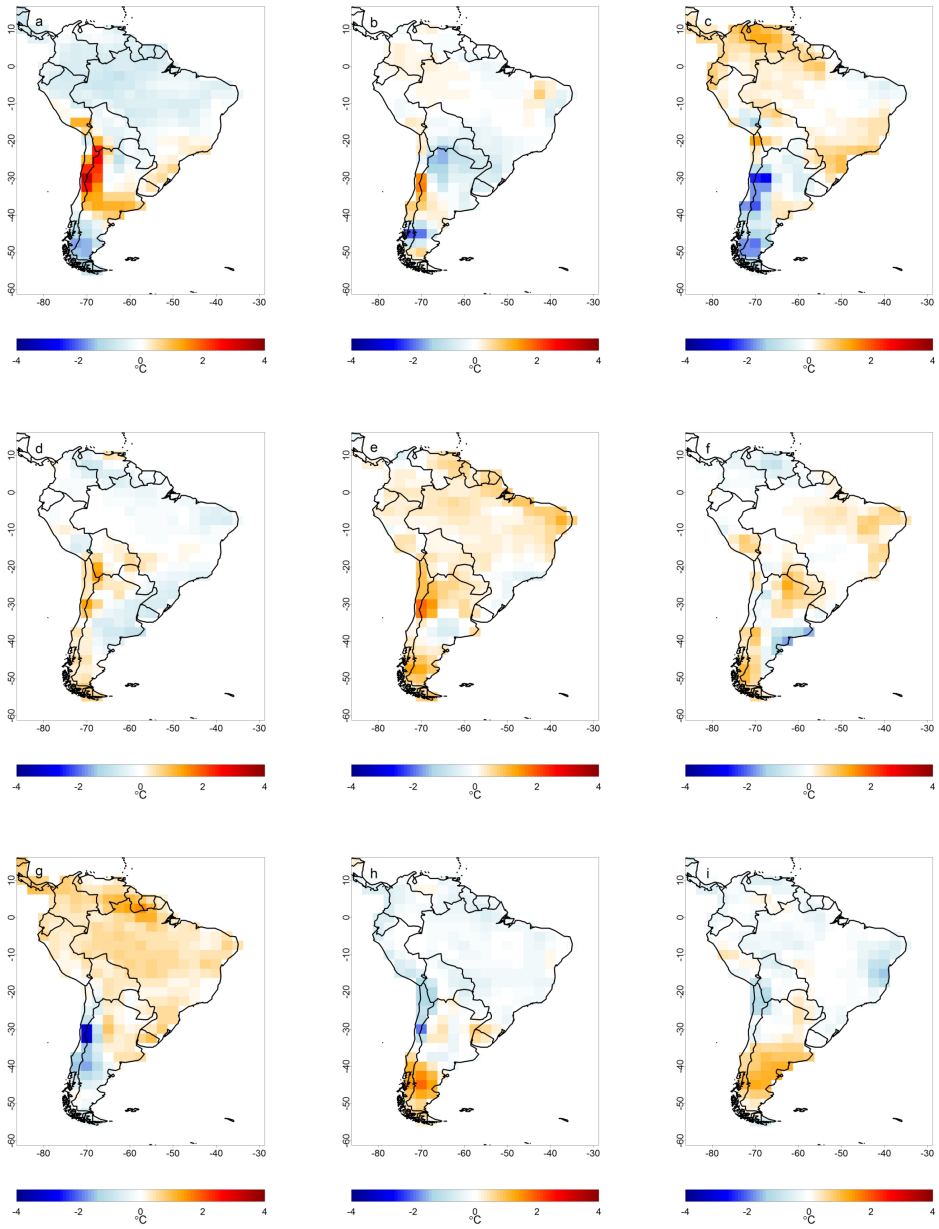


Figure 3.11: Temperature anomalies, relative to DJF 2000-01 to 2008-09 climatology, in South America for DJF season in (a) 2000-01, (b) 2001-02, (c) 2002-03, (d) 2003-04, (e) 2004-05, (f) 2005-06, (g) 2006-07, (h) 2007-08 and (i) 2008-09.

from sea level to almost 3000 metres. Secondly, the Caatinga zone, which is a dry forest region characterised by a semi-arid climate, located on the North/North East coast (see Fig. 3.6c). Figure 3.13 illustrates the mean annual cycle of DIR, precipitation and temperature for the South East Atlantic Rainforest and Caatinga, calculated over the period 2001-2009. In the South East Atlantic Rainforest, DIR peaks in March (Fig. 3.13.1a),

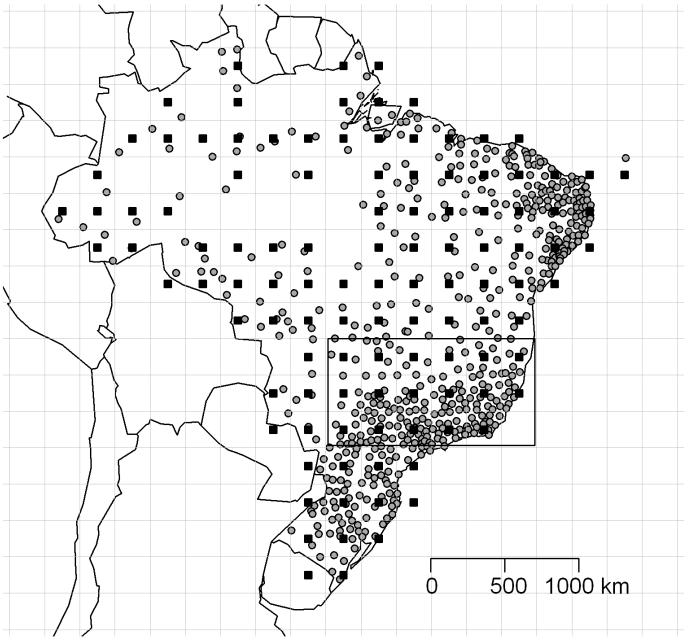


Figure 3.12: Centroids of microregions in Brazil (circles) and the 2.5° × 2.5° climate grid (squares). Large box indicates approximate location of South East region for which the modelling framework will be developed (see Chapter 5).

while precipitation peaks 2 months earlier in January (Fig. 3.13.1b). Temperature peaks in February and the coolest month is July (Fig. 3.13.1b). In Caatinga, DIR peaks in April (Fig. 3.13.2a) and rainfall peaks in March (Fig. 3.13.2b). While it rains less in Caatinga than in the South Atlantic Rainforest, the average temperature is several degrees higher (Fig. 3.13.2b).

Figure 3.14 shows scatter plots of precipitation/temperature and DIR in the selected zones for every month (2001-2009) and microregion contained with the zone. There is a weak positive association between precipitation and dengue incidence in the South East Atlantic Rainforest (Fig. 3.14a). As indicated by the shape of the local polynomial regression curve, a linear association is less obvious in the Caatinga zone (Fig. 3.14b). There is a positive association between temperature and dengue incidence in the South East Atlantic Rainforest (Fig. 3.14c). However, the association in the Caatinga zone is weaker (Fig. 3.14d).

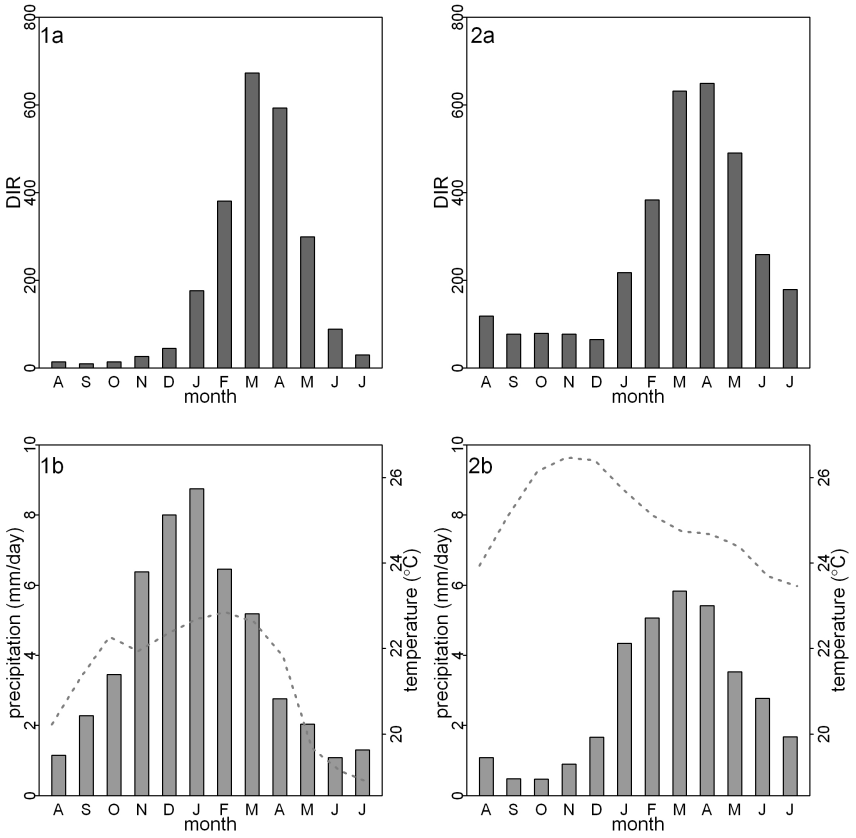


Figure 3.13: Mean annual cycle of (a) DIR and (b) precipitation and temperature (dashed line) in South East Atlantic Rainforest (column 1) and Caatinga in North East (column 2) for period 2001-2009.

3.4.4 Dengue and ENSO

A time series of the Oceanic Niño Index (ONI), defined as the 3-month running mean of SST anomalies in the Niño 3.4 region (120°W-170°W and 5°S- 5°N), based on the 1971-2000 base period, was obtained from the NOAA Climate Prediction Center (CPC)[9]. Using this index, the CPC define ENSO events when SST anomalies are $\geq +0.5$ for five consecutive months for warm (El Niño) events and ≤ -0.5 for cold (La Niña) events. During the study period of interest, a weak La Niña occurred in 2000-01, followed by a moderate El Niño in 2002-03. Weak El Niño events occurred in 2004-05 and 2006-07. A moderate La Niña occurred in 2007-08, followed by a strong El Niño in 2009-10 (see Fig. 3.15). In this study the ONI will be used as a continuous variable without reference to any threshold value for determining an El Niño or La Niña event.

[9] http://www.cpc.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml [accessed 15 May 2010]

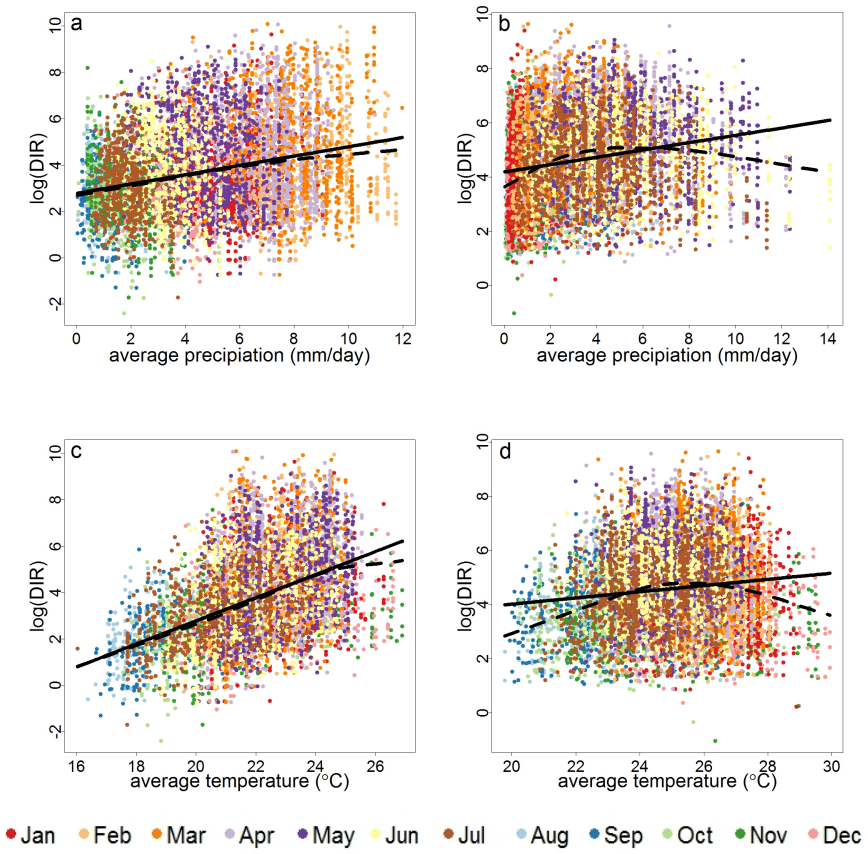


Figure 3.14: Scatter plot between DIR and precipitation in (a) South East Atlantic Rainforest, (b) Caatinga and temperature in (c) South East Atlantic Rainforest and (d) Caatinga. Climate variables averaged over 3 months previous to dengue month. Solid curve - linear model fit, dashed curve - local polynomial regression fit. Note points stratified by calendar month for DIR.

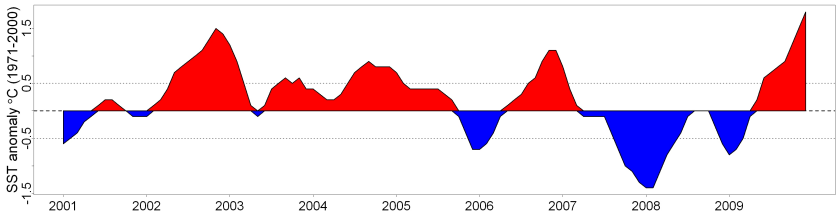


Figure 3.15: Oceanic Niño Index January 2001 - December 2009.

Figure 3.16 presents correlation maps between the Oceanic Niño index (ONI) and precipitation/temperature for the DJF season (2000-2009), with the ONI lagged from zero to five months. There is a negative association between ONI and precipitation over North

and North West Brazil and a positive association over the rest of the country. These associations change little with lag, particularly in the South East region. Temperature and ONI are positively associated across the whole of Brazil with a stronger association in the North and West and an area covering the South East region. The association appears to be consistent as the lag increases to five months. According to these data, during an El Niño event (positive ONI) drier and warmer conditions are expected in the North and West of Brazil and wetter and warmer conditions in parts of North East Brazil and South East Brazil in the DJF season. During a La Niña event, below average temperatures are expected across much of Brazil and above average rainfall in the Northern regions. The consistent signal with lag shows the ONI may be used to predict such tendencies. This is encouraging for predictive purposes. The robustness of these relationships is confirmed by comparing Figure 3.16 with Figure 3.17, where a 30 year period (1979-2009) is used to calculate the correlations. One notable difference is that the negative association between ONI and precipitation extend across a larger area of North Brazil when using the 30 year dataset. However, in general the key features described above appear to be similar when using both datasets. This is promising for interpreting climate associations using the limited dataset (2000-2009).

Figure 3.18 shows the association between the ONI and precipitation and temperature in the selected zones for the time period 2001-2009, with ONI lagged by four months. As expected from the previous analysis, there is a weak positive association between ONI (lag 4) and 3-month-average precipitation in South East Atlantic Rainforest (Fig. 3.18a) and a weak negative association in Caatinga (Fig. 3.18b). There is a weak positive association between ONI (lag 4) and 3-month-average temperature in both zones.

Figure 3.19 shows the association between the ONI and DIR in the selected zones for the time period 2001-2009, with ONI lagged by 2 months (simultaneous to climate variables). Various time lags between ONI and DIR were inspected. There was a slight negative relationship between ONI and DIR at lags ranging from 2 months to 6 months previous. It is interesting to observe a negative association between ONI and DIR, particularly in the South East region where ONI is positively associated with precipitation and temperature which are, in turn, positively associated with DIR. This will be investigated further in a modelling context in Chapter 4.

Figure 3.20 illustrates the time evolution of 3-month running average DIR, precipitation

and temperature anomalies and the ONI from January 2001 - December 2009 aggregated to the zone level, for the South East Atlantic Rainforest and Caatinga zones. Above average DIR in the first part of 2002 in the South East Atlantic Rainforest was accompanied by near average precipitation, temperature and Pacific SSTs (see Fig. 3.20.1). In the Caatinga, above average DIR in 2002 was preceded by above average precipitation (see Fig. 3.20.2). The moderate El Niño in 2002-03 coincided with above average precipitation and temperature, yet a below average DIR in South East Atlantic Rainforest. Below average DIR in 2004 in the South East Atlantic Rainforest was accompanied by below average temperatures, while below average DIR in Caatinga for the same period was preceded by well above average precipitation. The moderate La Niña in 2007-08 was accompanied by above average precipitation, below average temperature and the 2008 dengue epidemic in both zones.

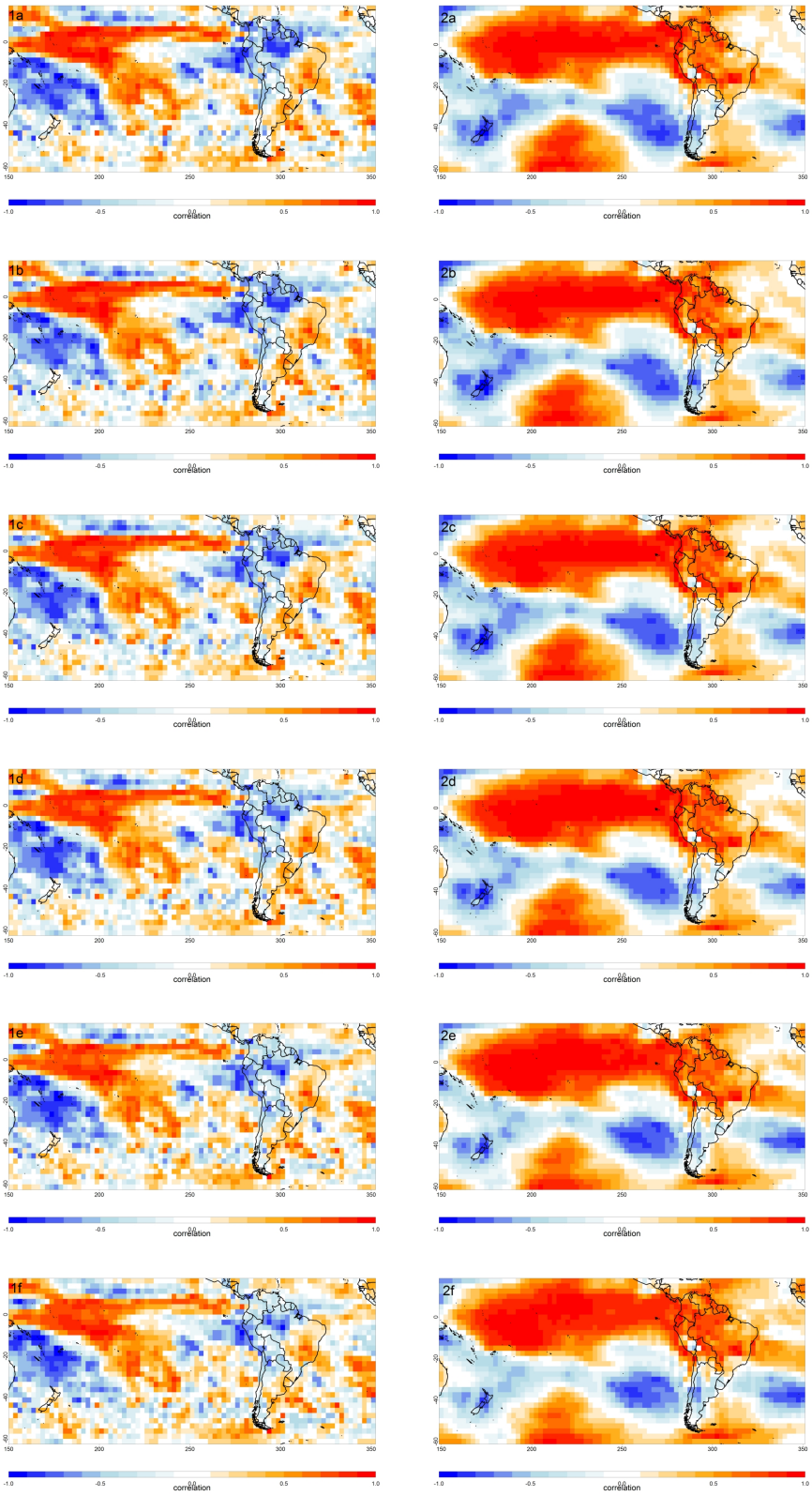


Figure 3.16: Correlation between gridded DJF average precipitation rate (column 1) and surface temperature (column 2) from 2000-01 to 2008-09 with the ONI lagged by (a) 0 months (DJF), (b) 1 month (NDJ), (c) 2 months (OND), (d) 3 months (SON), (e) 4 month (ASO) and (f) 5 months (JAS).

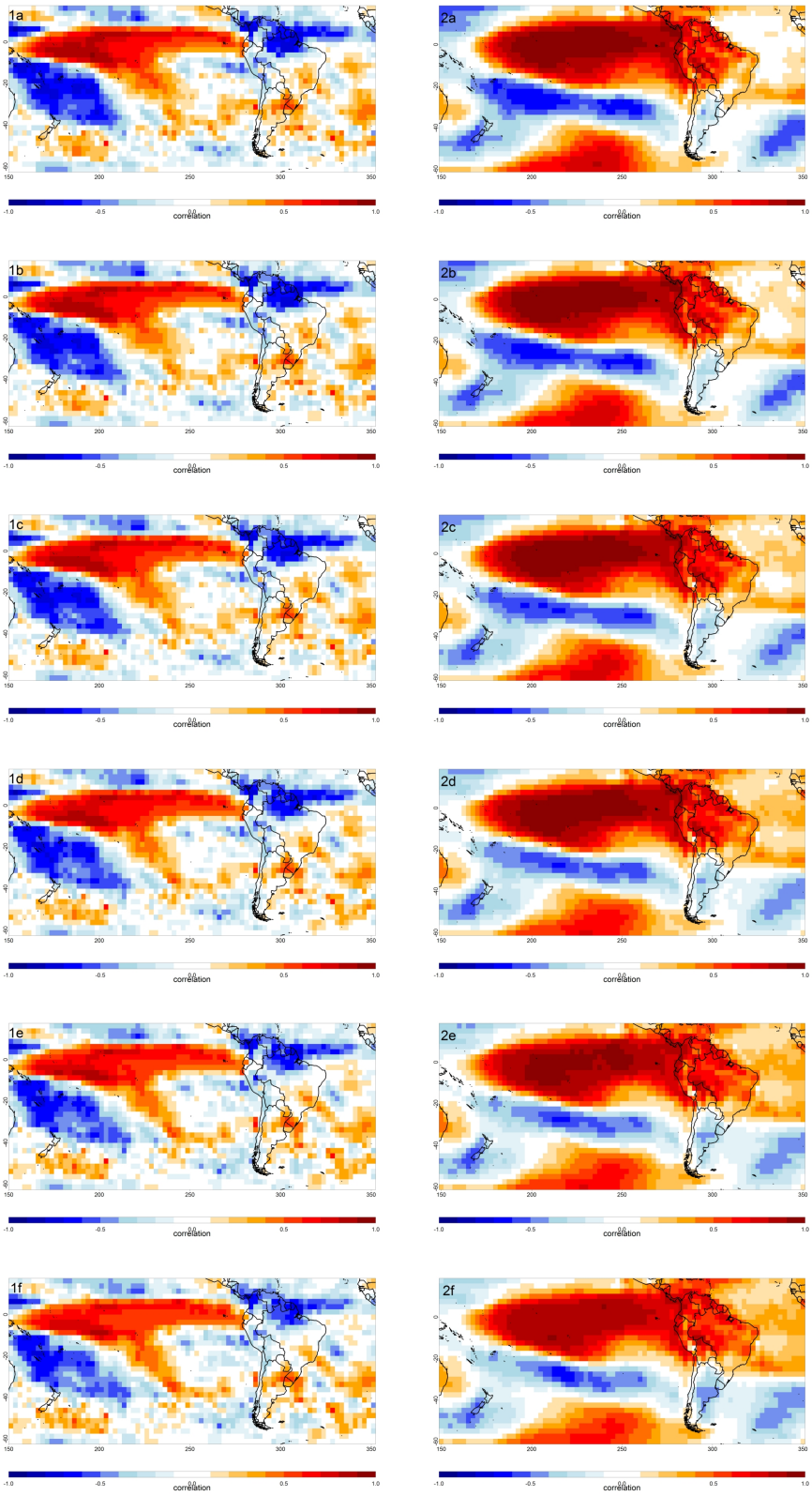


Figure 3.17: Correlation between gridded DJF average precipitation rate (column 1) and surface temperature (column 2) from 1979-80 to 2008-09 with the ONI lagged by (a) 0 months (DJF), (b) 1 month (NDJ), (c) 2 months (OND), (d) 3 months (SON), (e) 4 month (ASO) and (f) 5 months (JAS).

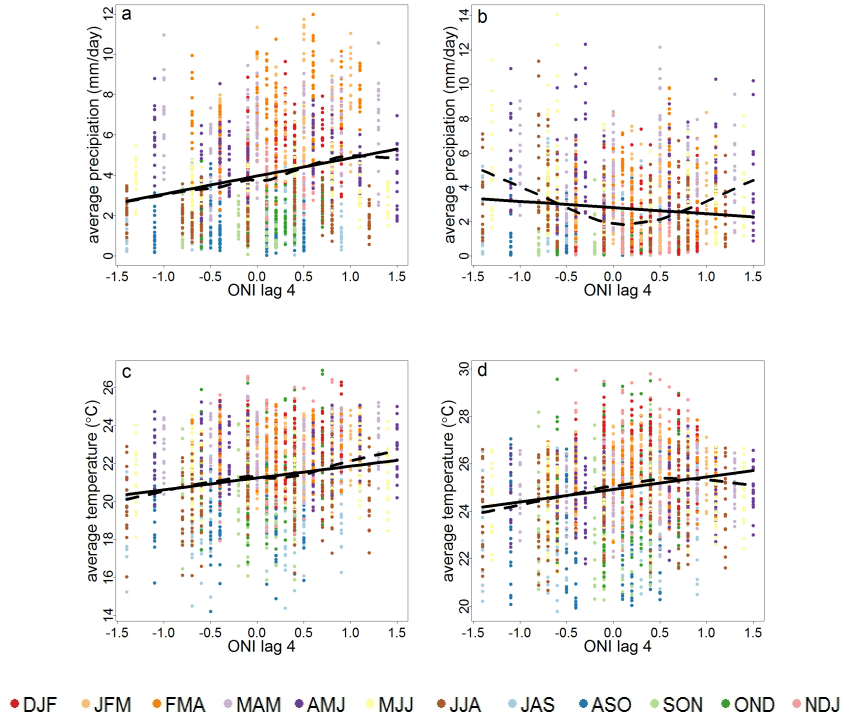


Figure 3.18: Scatter plot between ONI and precipitation, in (a) South East Atlantic Rainforest, (b) Caatinga, and temperature in (c) South East Atlantic Rainforest, (d) Caatinga. ONI lagged 4 months previous to climate variables. Solid curve - linear model fit, dashed curve - local polynomial regression fit. Note points stratified by 3-month seasons for climate variables.

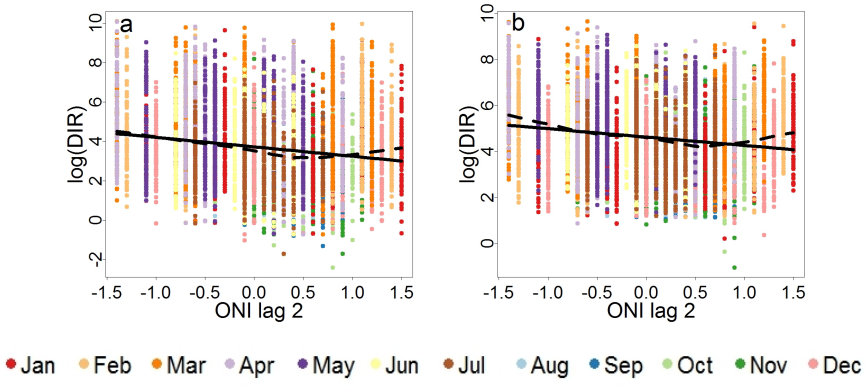


Figure 3.19: Scatter plot between ONI and DIR, in (a) South East Atlantic Rainforest, (b) Caatinga. ONI lagged 2 months previous to DIR. Solid curve - linear model fit, dashed curve - local polynomial regression fit. Note points stratified by calendar month for DIR.

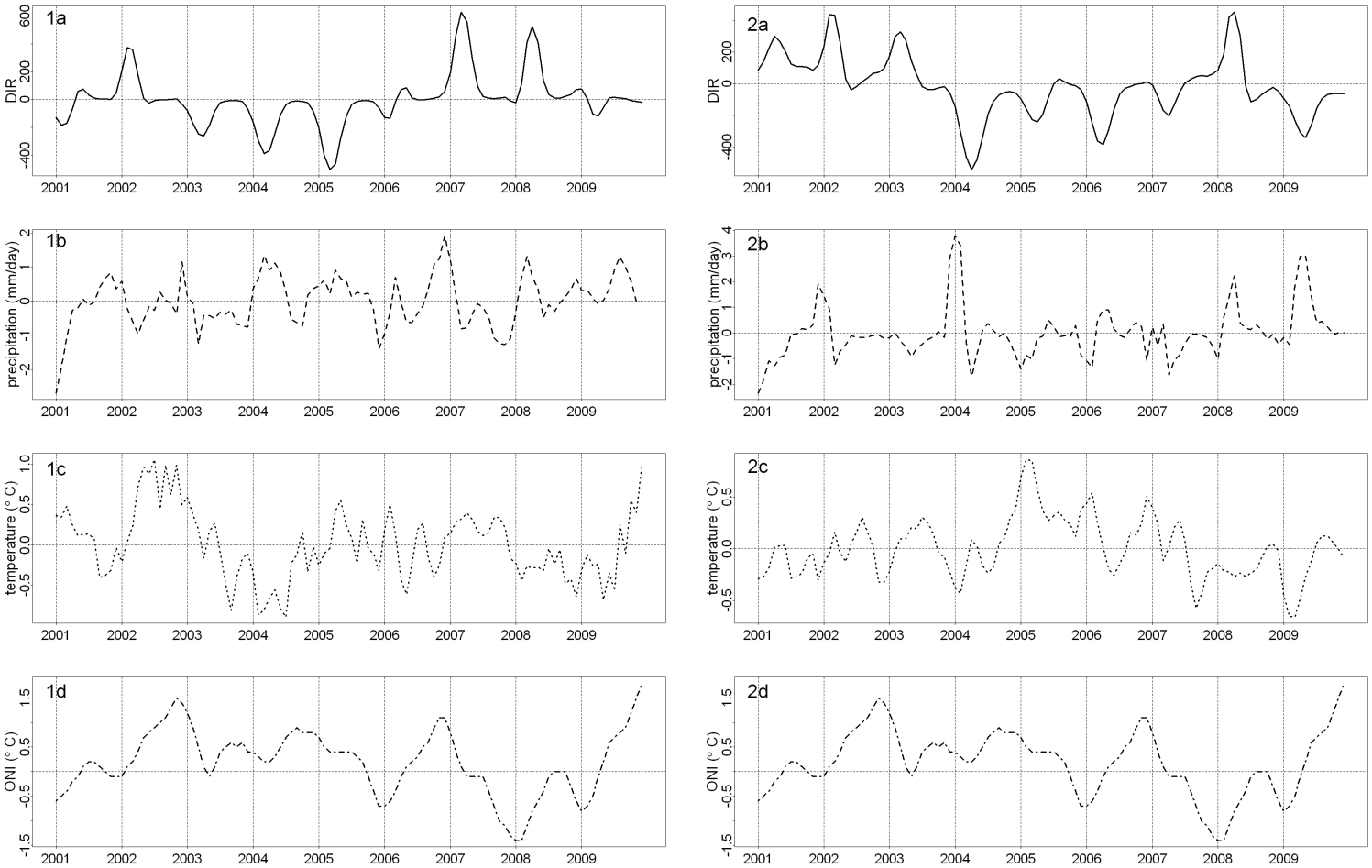


Figure 3.20: Time series of (a) 3-month running average DIR anomalies, (b) 3-month running average precipitation anomalies, (c) 3-month running average temperature anomalies and ONI from January 2001 - December 2009 for South East Atlantic Rainforest zone (column 1) and Caatinga zone (column 2).

3.5 Summary

In this chapter, the datasets collated for use in this study have been presented and explored. Dengue incidence in Brazil has been shown here to be associated with many different factors, e.g. altitude, urban population, geographical zone, temperature, precipitation and ONI. In order to understand and quantify the contribution of each component to the relative risk of dengue, a more complex model is needed to incorporate all possible confounding factors. In the next chapter, a modelling framework is proposed to model spatio-temporal variations in dengue relative risk in Brazil based on a combination of climatic and non-climatic explanatory variables.

Chapter 4

Model framework

4.1 Introduction

The aim of this chapter is to identify and test an appropriate probability model for dengue counts and to select suitable climate variables and other explanatory variables. This model provides the basis for the spatio-temporal hierarchical models that will be developed in Chapter 5. The predictive power of these models will be assessed in Chapter 6. In this chapter, the model and variable selection process is conducted within a generalised linear model (GLM) framework.

4.2 Generalised linear model framework

Count data can be modelled using the GLM framework (Nelder and Wedderburn, 1972). GLMs extend linear regression models to accommodate non-normal response distributions (Venables and Ripley, 2002). A GLM consists of two components; a probability distribution for the response variable within the exponential family of distributions and a monotonic differentiable link function that relates the mean to a linear predictor, involving any explanatory variables (covariates). The following provides a brief outline of the conceptual framework. For a detailed theoretical account of GLMs, see McCullagh and Nelder (1989).

GLMs describe the dependence of a random variable y_i $(i = 1, \ldots, n)$ on a set of explana-

tory variables $\mathbf{x}_i$. The distribution of the response variable y_i is considered to be in the exponential family, with probability density function

$$p(y_i, \lambda_i, \varphi) = \exp\left\{\frac{y_i\lambda_i - b(\lambda_i)}{\varphi} + c(y_i; \varphi)\right\}. \tag{4.1}$$

λ_i is the canonical parameter that depends on the explanatory variables via a linear predictor $\boldsymbol{\theta}\mathbf{x}_i'$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is a vector of p unknown parameters and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ is a vector of p explanatory variables. φ is a dispersion parameter that is possibly known. For example, $\varphi = 1$ in discrete and count models such as the binomial or Poisson distribution (Hilbe, 2007). The functions $b(\cdot)$ and $c(\cdot)$ are known and determine which member of the family is used, e.g. normal, binomial or Poisson distribution (Zeileis et al., 2008). The mean and variance of y_i are given by

$$E[y_i] = \mu_i = b'(\lambda_i)$$
$$Var[y_i] = \sigma_i^2 = \varphi\, b''(\lambda_i),$$

where $b'(\lambda_i)$ and $b''(\lambda_i)$ are the first and second derivatives of $b(\lambda_i)$. Therefore, up to a dispersion parameter φ, the distribution of y_i is determined by its mean. The variance of y_i is proportional to the variance function $V(\mu_i) = b''(\lambda_i)$ which only depends on the canonical parameter.

The dependence of the mean $E[y_i] = \mu_i$ on the explanatory variables $\mathbf{x}_i$ is specified via

$$g(\mu_i) = \boldsymbol{\theta}\mathbf{x}_i',$$

where $g(\cdot)$ is a specified link function that allows a non-linear relationship between the mean μ_i of the response and the linear function of the explanatory variables $\boldsymbol{\theta}\mathbf{x}_i'$. Therefore, $\mu_i = g^{-1}(\boldsymbol{\theta}\mathbf{x}_i')$ is an estimate of the mean of the ith observation, obtained from an estimate of the parameter vector $\boldsymbol{\theta}$. Maximum likelihood methods can be used to estimate the unknown parameters $\boldsymbol{\theta}$. This can be implemented through a technique known as iterative re-weighted least squares (IRLS, see Appendix A). IRLS algorithms are available in standard statistical computing software such as R (R Development Core Team, 2009). R provides a flexible implementation of the GLM framework in the function `glm()` (Chambers and Hastie, 1992) contained in the **stats** package.

4.2.1 Poisson model

Let us consider a basic model framework for dengue counts in the 558 microregions of Brazil from January 2001 - December 2009 (108 months). The simplest distribution used for modelling such count data is the Poisson distribution with probability density function

$$p(y_i;\mu_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}, \quad y_i = 0,1,2,\ldots, \quad \text{and} \quad \mu_i > 0.$$

This gives the probability that a particular y_i value is observed for a given mean μ_i, where μ_i is a function of the covariates. The Poisson distribution is a special case of the exponential family (Eqn. 4.1) with canonical parameter $\lambda_i = \log \mu_i$. The canonical link function is $g(\mu_i) = \log \mu_i$, resulting in a log-linear relationship between the mean and linear predictor (Zeileis et al., 2008). With φ=1, one obtains $V(\mu_i) = \mu_i$, a defining property of the Poisson distribution that the variance equals the mean. With a logarithmic link the mean is multiplicative:

$$\mu_i = \exp(\boldsymbol{\theta}\mathbf{x}_i').$$

Let y_{st} denote dengue counts in space s and time t, where $s = 1,\ldots,S$, with $S = 558$ and $t = 1,\ldots,T$, with $T = 108$ (total number of observations $n = S \times T = 60264$). Consider the counts to be Poisson distributed:

$$\begin{aligned} y_{st} &\sim \text{Pois}(\mu_{st}) \\ \log \mu_{st} &= \log e_{st} + \log \rho_{st} \\ \rho_{st} &= \prod_{j=1}^{p} \exp(\theta_j x_{jst}), \end{aligned}$$

where for each spatial location and time period, counts of dengue cases y_{st} follow a Poisson distribution with mean μ_{st} which is equal to the expected number of cases e_{st} (see Eqn. 3.2, Section 3.2.3, Chapter 3) multiplied by the unknown relative dengue risk ρ_{st} for a given microregion s and time t. The likelihood for the data is then

$$L = \prod_{s=1}^{S}\prod_{t=1}^{T} \frac{\exp(-e_{st}\rho_{st})\ (e_{st}\rho_{st})^{y_{st}}}{y_{st}!}.$$

The log-likelihood function is given by

$$l = \log L = \sum_{s=1}^{S}\sum_{t=1}^{T}[y_{st}(\log e_{st} + \log \rho_{st}) - e_{st}\rho_{st} - \log y_{st}!].$$

By including the expected number of cases e_{st} in each region as an offset, the most suitable estimate for the relative risk ρ_{st} is sought via a linear combination of climate covariates (temperature and precipitation), a large scale climate driver (ONI) and non-climate confounding factors, i.e. cartographic and demographic covariates, that might explain variations in dengue risk.

4.2.2 Negative binomial model

Although Poisson models are a popular choice for the analysis of count data, it is well established that observed count data, e.g. disease cases, often display substantial extra-Poisson variation, or overdispersion (Lawless, 1987). In other words, $Var[y_i] > E[y_i]$; the variance of the response variable exceeds the mean.

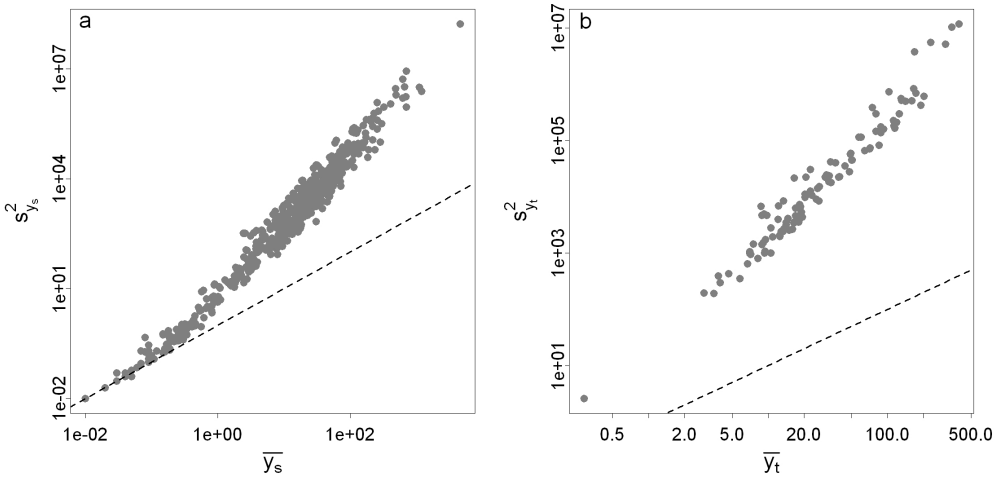


Figure 4.1: Mean and variance of dengue counts (a) for each microregion, $s = 1, \ldots, 558$, over all months and (b) for each month, $t = 1, \ldots, 108$, over all microregions. Dashed line - mean=variance.

An indication of the magnitude of overdispersion (or underdispersion) can be obtained simply by comparing the sample mean ($\bar{y} = 54$) and variance ($s_y^2 = 432213$) of the dependent count variable (Cameron and Trivedi, 1998). Figure 4.1 illustrates that the variance increases much more rapidly than the mean when comparing the sample mean $\bar{y}_s$ and variance $s_{y_s}^2$ for each microregion over the time period (108 months) (see Fig 4.1a)

$$\bar{y}_s = \frac{1}{T} \sum_{t=1}^{T} y_{st}$$

$$s^2_{y_s} = \frac{1}{T-1}\sum_{t=1}^{T}(y_{st} - \bar{y}_{st})^2$$

and the sample mean $\bar{y}_t$ and variance $s^2_{y_t}$ for each time period over all 558 microregions (see Fig 4.1b)

$$\bar{y}_t = \frac{1}{S}\sum_{s=1}^{S} y_{st}$$

$$s^2_{y_t} = \frac{1}{S-1}\sum_{s=1}^{S}(y_{st} - \bar{y}_{st})^2.$$

A Poisson model of an overdispersed variable will lead to an underestimate of the standard errors that will overstate the significance of the parameter estimates, resulting in misleading inference. For example, a variable may appear to be statistically significant when in fact it is not. In Chapter 5, explicit allowances for this overdispersion are considered via the inclusion of appropriate random effects. For model selection purposes within this chapter, overdispersion is addressed by using the negative binomial distribution; a model often used to account for overdispersion in counts (Cameron and Trivedi, 1998; Hilbe, 2007). The negative binomial distribution is given by

$$p(y_i; \mu_i, \kappa) = \frac{\Gamma(y_i + \kappa)}{\Gamma(\kappa) y_i!}\frac{\mu_i^{y_i}\kappa^{\kappa}}{(\mu_i + \kappa)^{y_i+\kappa}}, \tag{4.2}$$

with mean μ_i, scale parameter κ and variance function $V(\mu_i) = \mu_i + \mu_i^2/\kappa$. For fixed κ, Equation 4.2 is a special case of the exponential family (Eqn. 4.1) with canonical parameter $\lambda_i = \log(\frac{\mu_i}{\mu_i+\kappa})$. If κ is not known, but is to be estimated from the data, the negative binomial is no longer a special case of the exponential family. In this case, maximum likelihood can be used, leading to estimates of both $\boldsymbol{\theta}$ and κ, using the function `glm.nb()` from the **MASS** package (Venables and Ripley, 2002) in R. Here, the log-link $g(\mu_i) = \log \mu_i$ is used to allow comparison of point estimates to the Poisson model. As discussed by Dean and Lawless (1989), the negative binomial can allow for overdispersion caused by omitting important variables from the model. As $\kappa \to \infty$ the negative binomial reduces to the Poisson model (i.e. $V(\mu_i) \to \mu_i$). The smaller the value of κ, the more variability there is in the data over and above that associated with the mean μ_i. Using the negative binomial, the modelling framework for dengue counts is:

$$\begin{aligned} y_{st} &\sim \text{NegBin}(\mu_{st}, \kappa) \\ \log \mu_{st} &= \log e_{st} + \log \rho_{st} \\ \rho_{st} &= \prod_{j=1}^{p} \exp(\theta_j x_{jst}). \end{aligned}$$

As in Section 4.2.1, y_{st} is the dengue count, μ_{st} is the corresponding mean dengue count, κ is the scale parameter and the expected cases e_{st} are treated as an offset in the model.

4.3 Overdispersion

Overdispersion to some degree is inherent to the vast majority of count data (Hilbe, 2007). Therefore, an interesting question concerns the amount of overdispersion in a particular model - is it statistically sufficient to require a model other than Poisson? A decision about whether the Poisson form is appropriate can be based on one of several statistics. The advantage of using the maximum likelihood method is that a simple likelihood ratio test may be employed to assess the adequacy of the negative binomial over the Poisson, as the negative binomial reduces to the Poisson as $\kappa \to \infty$. In other words, the Poisson is a special case of the negative binomial with scale parameter $\kappa = \infty$ or 'overdispersion' parameter $\kappa^{-1} = 0$. For testing a Poisson model against a negative binomial model, the hypotheses may be stated as: $H_0 : \kappa^{-1} = 0$ against $H_1 : \kappa^{-1} > 0$. The likelihood ratio test-statistic is:

$$T = -2(l_1(\hat{\boldsymbol{\mu}}; \mathbf{y}) - l_0(\hat{\boldsymbol{\mu}}; \mathbf{y})),$$

where $l_1(\hat{\boldsymbol{\mu}}; \mathbf{y})$ is the log-likelihood for the Poisson model and $l_0(\hat{\boldsymbol{\mu}}; \mathbf{y})$ is the log-likelihood for the negative binomial model, fitted to the same data. The likelihood ratio test-statistic has a non-standard distribution since the negative binomial 'overdispersion' parameter κ^{-1} is restricted to be positive. The asymptotic distribution of the likelihood ratio test-statistic has probability mass of 0.5 at zero and a $0.5\chi^2_1$ distribution for positive values (see Cameron and Trivedi, 1998, p.78), where χ^2_k is a chi-squared distribution with k degrees of freedom ($k = 1$ in a test of the Poisson model against the negative binomial model). Therefore, to test the null hypothesis at the significance level of $\varpi = 0.05$, the critical value of the χ^2_1 distribution with significance level $2\varpi = 0.10$ is used (i.e. a one-sided test). Therefore, H_0 is rejected if $T > \chi^2_{0.90,1} = 2.7$.

4.4 Goodness-of-fit

There are several goodness-of-fit criteria available to determine how well a specified model fits the data. In conventional generalized linear modelling with fixed effects, the deviance

D is an important measure (McCullagh and Nelder, 1989). This measure of model adequacy compares a fitted model to a saturated model, i.e. an exact fit in which the fitted values are equal to the observed data. It is based on the difference between the log-likelihood of the data under both models:

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2\left[l(\mathbf{y};\mathbf{y}) - l(\hat{\boldsymbol{\mu}};\mathbf{y})\right],$$

where $l(\mathbf{y};\mathbf{y})$ is the log-likelihood for a saturated model. Maximising the likelihood of the fitted model $l(\hat{\boldsymbol{\mu}};\mathbf{y})$ is equivalent to minimising $D(\mathbf{y};\hat{\boldsymbol{\mu}})$ with respect to the fitted values $\hat{\boldsymbol{\mu}}$.

For an adequate model, the scaled deviance D/φ, has an asymptotic chi-squared distribution with $n-p$ degrees of freedom. For count data probability models, the dispersion parameter, $\varphi = 1$ (see Section 4.2). Thus for these models we expect $D \sim \chi^2_{n-p}$. The expected value of a chi-squared distribution is equal to its degrees of freedom, $E[\chi^2_{n-p}] = n-p$. If the model is correctly specified, the ratio of the deviance to $n-p$ should be close to unity (Crawley, 2002). Therefore, if the values for D are close to the degrees of freedom, the model may be considered as adequate. If the ratio is substantially greater than 1, i.e. $D > n-p$, then the data are overdispersed and alternative measures to accommodate overdispersion are recommended (e.g. specification of a negative binomial model).

When using Normal theory linear models, R^2 (the proportion of variation in the response variable that is account for by the model) is typically used as a measure of model adequacy. In the GLM context, a number of alternative measures have been proposed (Cameron and Trivedi, 1998). The deviance, defined above, can be thought of as the GLM generalisation of the sum of squares. Note that for the Normal distribution, the deviance is just the residual sum of squares. Cameron and Windmeijer (1996) propose a pseudo-R^2 based on decomposition of the deviance:

$$D(\mathbf{y};\bar{\mathbf{y}}) = D(\mathbf{y};\hat{\boldsymbol{\mu}}) + D(\hat{\boldsymbol{\mu}};\bar{\mathbf{y}}),$$

where $D(\mathbf{y};\bar{\mathbf{y}})$ is the deviance in the intercept-only model (i.e. null model), $D(\mathbf{y};\hat{\boldsymbol{\mu}})$ is the deviance in the fitted model, and $D(\hat{\boldsymbol{\mu}};\bar{\mathbf{y}})$ is the explained deviance. Then

$$R^2_D = 1 - \frac{D(\mathbf{y};\hat{\boldsymbol{\mu}})}{D(\mathbf{y};\bar{\mathbf{y}})},$$

which measures the reduction in the deviance due to inclusion of explanatory variables.

This equals $D(\hat{\boldsymbol{\mu}}; \bar{\mathbf{y}})/D(\mathbf{y}; \bar{\mathbf{y}})$ and can be interpreted as the explained deviance. R_D^2 lies between 0 and 1, with values closer to 1 implying a better fit.

Several likelihood measures have been proposed in the statistical literature to compare the performance of alternative models. Akaike's information criterion (AIC) (Akaike, 1974) is a measure of goodness-of-fit of an estimated statistical model, which not only rewards goodness-of-fit but also includes a penalty that discourages overfitting:

$$AIC = -2l(\hat{\boldsymbol{\mu}}; \mathbf{y}) + 2p$$

where $l(\hat{\boldsymbol{\mu}}; \mathbf{y})$ is the log-likelihood of the fitted model and p is the number of parameters. The second term acts as a penalty for over parameterisation of the model. The smaller the AIC, the better the model is. The AIC is widely used for fixed effect models and is the basis of the deviance information criterion (DIC) which will be used in Chapter 5.

Another variant that is commonly used as a model choice criterion is the Bayesian information criterion (BIC) (Schwarz, 1978). It is defined as:

$$BIC = -2l(\hat{\boldsymbol{\mu}}; \mathbf{y}) + p \log n$$

where n is the total number of data points. Similarly, the smaller the BIC, the better the model.

4.5 Choice of distribution

In order to select the most appropriate distribution to model dengue counts in Brazil from January 2001 - December 2009, both Poisson and negative binomial GLMs were fitted to the dengue data with climate and non-climate covariates, factors and interactions combined in the linear predictor (further details provided in section 4.6). A summary of test results and information criteria is presented in Table 4.1. The likelihood ratio statistic T is highly significant (p-value $= 2 \times 10^{-16}$) with $T >> 2.7 = \chi^2_{0.90,1}$ (see Table 4.1). This suggests that there is strong evidence to reject the null hypothesis that the Poisson model is adequate for this data. For a correctly specified model, the deviance divided by the degrees of freedom should be close to unity. Fitting a Poisson GLM results in a residual deviance more than one hundred times larger than the residual degrees of freedom ($D/(n-p) = 118.3$), implying that the data are overdispersed.

Table 4.1: Likelihood ratio test statistic (T), degrees of freedom ($n-p$), Deviance (D), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for a combined model fit using Poisson and negative binomial GLMs.

Test/Criteria	Poisson	Negative Binomial
T	-	6901127
$n-p$	60142	60141
D	7114161	60520
AIC	7275640	374515
BIC	7276739	375612

Apparent dispersion can occur when the data include outliers (see Hilbe, 2007), along with the omission of important explanatory variables or an insufficient number of interaction terms. The Poisson GLM was re-fitted by removing the three most highly influential points (evident in the upper right hand corner of Fig. 4.4 and 4.5a). The quotient of deviance and degrees of freedom remained much greater than 1 ($D/(n-p) = 109.7$). This implies that real overdispersion exists and an alternative model or technique should be adopted to deal with this. By comparison, the deviance is close to the degrees of freedom for the negative binomial GLM ($D/(n-p) = 1.006$), indicating that this model is much better specified. There is a large reduction in both the AIC and BIC when fitting a negative binomial GLM. These criteria indicate that the negative binomial provides a better fit. Figure 4.2 gives a comparison of the mean-variance relationship for the Poisson model where the mean and variance function are equal to μ_i, estimated from the Poisson GLM fit, and the negative binomial model where the variance function is given by $\mu_i + \mu_i^2\kappa^{-1}$. μ_i and κ^{-1} were estimated from the negative binomial GLM fit with $\hat{\kappa}^{-1} = 3.18$.

Figure 4.3 illustrates kernel density estimates for the observed data y_{st}, randomly generated Poisson samples with mean $\hat{\mu}_{st}$ obtained from the Poisson GLM, $\hat{y}_{st} \sim \text{Pois}(\hat{\mu}_{st})$, and randomly generated negative binomial samples with mean $\hat{\mu}_{st}$ and scale parameter $\hat{\kappa}$ obtained from the negative binomial GLM, $\hat{y}_{st} \sim \text{NegBin}(\hat{\mu}_{st}, \hat{\kappa})$. The probability that the number of dengue cases for a given microregion and month will take the value of 20 ($\log(20) = 3$), for example, is much greater for the Poisson density estimates than that observed in the data. Conversely, the probability of observing any number of dengue cases is similar for both the data and the negative binomial density estimates. The re-

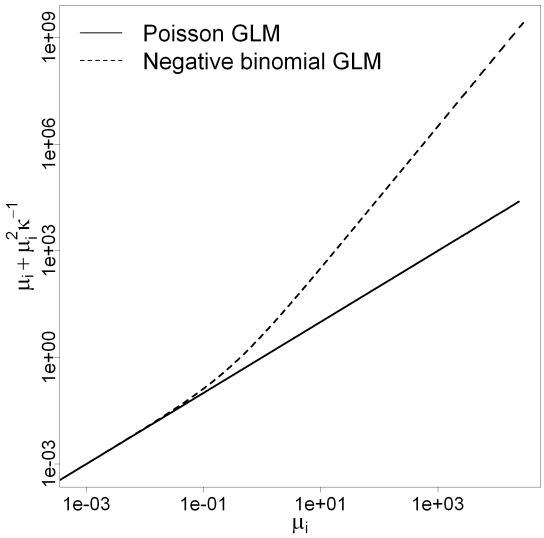


Figure 4.2: Comparison of variance in dengue counts for Poisson model (solid line) where $\kappa^{-1} = 0$ and negative binomial model where $\hat{\kappa}^{-1} = 3.18$ (dashed line). Note logarithmic axes.

sults thus far suggest that the negative binomial GLM provides a better fit than the Poisson GLM.

4.5.1 Residual analysis

Residual deviance is the GLM equivalent of the residual sum of squares ($\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$) and small independent residuals are desirable. The contribution of each observation (dengue count) to the residual deviance can be found by calculating the deviance residuals:

$$r_{(D)i} = sgn(y_i - \hat{\mu}_i)\sqrt{d_i} \tag{4.3}$$

where d_i is the contribution of the ith observation to the deviance and $\sum {r_{(D)i}}^2 = D$, the deviance (McCullagh and Nelder, 1989). For the Poisson model

$$d_i = 2\left[y_i \log(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i\right]$$

and for the negative binomial model with known κ,

$$d_i = 2\left[y_i \log(y_i/\hat{\mu}_i) - (y_i + \hat{\kappa}) \log\left(\frac{y_i + \hat{\kappa}}{\hat{\mu}_i + \hat{\kappa}}\right)\right]. \tag{4.4}$$

An overall test of adequacy of a linear model may be to see how close the residuals are to normality (Cameron and Trivedi, 1998). This can be done by a normal score plot,

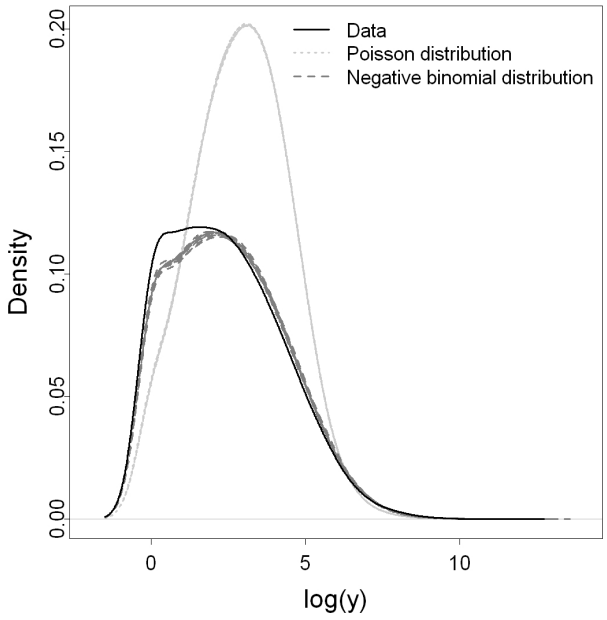


Figure 4.3: Kernel density estimates for dengue counts y_{st} (black line), 10 randomly generated Poisson samples with mean $\mu_{st} = \hat{\mu}_{st}$ estimated from the fitted Poisson GLM (light grey lines) and 10 randomly generated negative binomial samples with mean $\mu_{st} = \hat{\mu}_{st}$ and $\kappa = \hat{\kappa}$ estimated from the fitted negative binomial GLM (dark grey lines). Note logarithmic axes.

which orders the residuals from smallest to largest and plots them against theoretical values of the Normal distribution. If the residuals are exactly Normal, this produces a 45° straight line. Davison and Gigli (1989) advocate using normal scores plots with deviance residuals to check distributional assumptions. For a GLM, we do not expect the deviance residuals to be normally distributed (Ben and Yohai, 2004; Zuur et al., 2009), but we are interested in detecting outliers (Faraway, 2006). Figure 4.4 illustrates the extreme behaviour of the deviance residuals in the Poisson GLM, and the presence of outliers with deviance residuals exceeding 300 (see top right corner). Comparatively, the deviance residuals from the negative binomial are much smaller.

For all GLMs except the Gaussian, the variance function is non-constant. However, by using deviance residuals the variance function is scaled out (Faraway, 2006). Therefore, we would expect to see constant variance in a plot of deviance residuals against the fitted values at the linear predictor scale, provided the variance function is correct. A violation of non-constant variance is termed heteroskedasticity (i.e. errors are not the same across the range of fitted values). Figure 4.5a shows a fanning out of the deviance residual

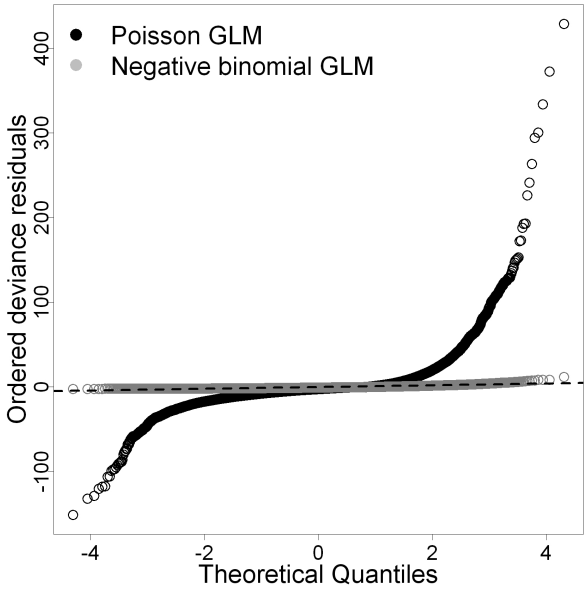


Figure 4.4: Ordered deviance residuals from Poisson (black) and negative binomial GLM (grey) in relation to theoretical quantiles for a Gaussian error distribution. Dotted line indicates 45° line.

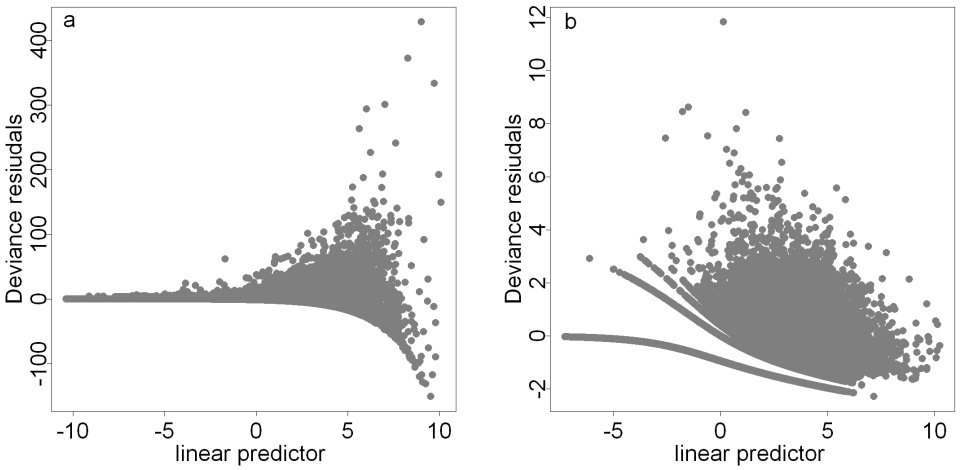


Figure 4.5: Plot of deviance residuals against fitted values at the linear predictor scale for (a) Poisson model and (b) negative binomial GLM.

pattern for the Poisson model, indicating such heteroskedasticity. The residual pattern for the negative binomial model does not exhibit any such pattern (see Fig. 4.5b).

Various tests and criteria including the likelihood ratio test, AIC, BIC and residual

analysis, suggest that the Poisson model is not appropriate to model dengue counts in Brazil. The negative binomial GLM appears to be a better model and hence will be employed in the following section for the selection of covariates.

4.6 Selection of covariates

In order to select which of the explanatory variables are important for modelling dengue counts in Brazil for the 108 month time period (January 2001 - December 2009), a negative binomial GLM was fitted to the dengue data. Along with approximate maximum likelihood estimates for the coefficients from a GLM, estimates of the standard error for each coefficient were obtained. From these, a p-value was calculated for the null hypothesis H_0 that the true value of each coefficient estimate is zero. A test known as the z-score was calculated as the ratio between the estimated coefficient and the estimated standard error. If the p-value is less than ϖ, the level of significance, H_0 is rejected and the coefficient estimate is considered to be significantly different from zero. Hence, the covariate is statistically significant at the $\varpi = 0.05$ level if p-value< 0.05.

The model selection was initiated with a maximal model based on all of the covariates described in the previous chapter, i.e. spatial covariates related to the urban environment, altitude, the annual cycle and interactions with geographical zone, observed monthly climate variables with associated time lags (0-3 months) and the Oceanic Niño Index (ONI) with time lags of up to 6 months. Quadratic terms related to these climate covariates were also considered in order to capture possible non-linear effects. Exploratory analyses were then carried out using different subsets of variables, to select an appropriate prediction model (e.g. examining model fit with and without climate information and with different interactions). These analyses were assisted by use of model selection algorithms based on the AIC stepwise regression.

At the 0.05 level of significance, precipitation and temperature covariates lag 1-3 were found to be statistically significant. These time lags are consistent with previous findings (e.g. Schreiber, 2001; Wu et al., 2007; Tipayamongkholgul et al., 2009; Johansson et al., 2009b). Rather than selecting a particular lag, or including all three lag separately, which could result in over-fitting, these variables were combined into 3-month average precipitation and temperature variables, lagged 2 months previous to the dengue month

of interest. As this model is intended to be used as an early warning system, temperature and precipitation would in practice be obtained from seasonal climate forecasting systems (e.g. EUROBRISA forecasting system, see Chapter 7 for more details). Such forecasts are typically issued as seasonal forecasts (e.g. December–February average) rather than monthly forecasts. The ONI lagged 2 months and 6 months prior to the dengue month of interest (or 4 months prior to the averaged temperature and precipitation effects) were favoured by the AIC stepwise model selection algorithm. As the relationship between the ONI and precipitation and temperature over parts of Brazil is consistent out to lags of 5 months (see Fig. 3.16), an ONI with lag 6 months prior to the dengue month of interest was adopted. This provides increased lead time which could be advantageous for a dengue early warning system. Observed ONI could be used with forecast temperature and precipitation, providing up to 5 months lead time. For example, if the Brazilian Ministry of Heath were interested in predicting dengue in March (when dengue peaks in many zones, e.g. Atlantic Rainforest and Cerrado zones), the required climate information would be December-February averaged precipitation and temperature and August-October ONI (see Fig. 4.6). As for non-linear terms in the climate variables, no strong evidence was found of the need to include non-linear terms in the climate covariates. This was confirmed by the lack of curvature in plots of residuals from a fitted model including only linear climate terms, against each climate covariate in turn. Therefore, the selected climate covariates were 3-month average precipitation and temperature, lagged 2 months and ONI, lagged 6 months.

A series of models of increasing complexity, from a global intercept to spatially and temporally varying covariates, interacted with geographical zone were tested (see Table 4.2). Various models were considered:

Null model	$\log \rho_{st} = \alpha$
Climate model	$\log \rho_{st} = \alpha + \sum_j \beta_j x_{jst}$
Non-climate model	$\log \rho_{st} = \alpha + \delta_{1t'(t)} + \delta_{2k(s)} + \delta_{3k(s)t'(t)} + \sum_j \gamma_j w_{jst}$
Combined model	$\log \rho_{st} = \alpha + \delta_{1t'(t)} + \delta_{2k(s)} + \delta_{3k(s)t'(t)} + \sum_j \gamma_j w_{jst} + \sum_j \beta_j x_{jst} + \sum_j \beta_{jk(s)} x_{jst}.$

The null model included only an intercept α in the relative risk $\log \rho_{st}$. The climate model included selected climate covariates x_{jst} with $j = 1, \ldots, 3$: 3-month average precipitation and temperature, lagged two months, and ONI, lagged 6 months. The non-climate model

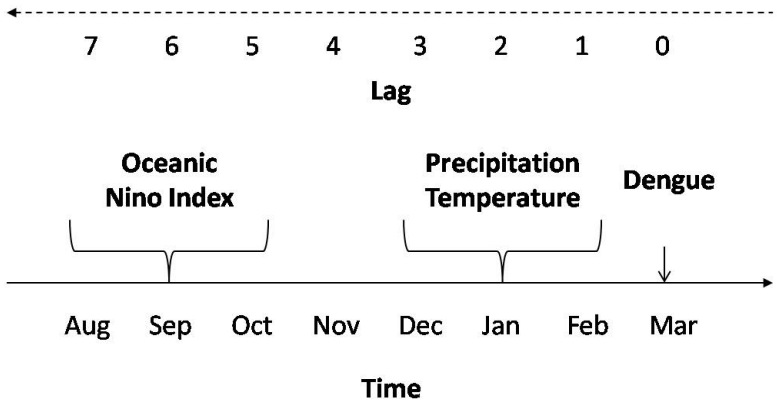


Figure 4.6: Schematic to show time lags between dengue month of interest (e.g. March), 3-month average precipitation and temperature lagged 2 months prior to dengue month (e.g. December-February) and ONI lagged 6 months prior to dengue month (e.g. August to October, 4 months prior to average precipitation and temperature).

included categorical variables: calendar month $\delta_{1t'(t)}$ with $t'(t) = 2, \ldots, 12$ (to account for the annual cycle), zone $\delta_{2k(s)}$ with $k(s) = 2, \ldots, 8$ (to account for differences between areas where climatic, geographical and ecological conditions are approximately homogeneous) and their interaction $\delta_{3k(s)t'(t)}$, to account for varying seasonal cycles between the geographic zones (see Fig. 3.8). Note that zone $k(s) = 1$ (Amazon Rainforest) and calendar month $t'(t) = 1$ (August) are set as the reference level and their effects are aliased in the intercept α. Demographic and geographic covariates $w_{jst}, j = 1, 2$ were altitude and population density. The combined model is an extension to the non-climate model with the addition of the three selected climate covariates and their interaction with geographical zone, providing a parameter estimate $\beta_{jk(s)}$ with $j = 1, \ldots 3$ and $k(s) = 2, \ldots, 8$ for each climate covariate. This represents the difference in climate effects between each zone in Brazil and the reference zone (Amazon Rainforest).

Results displayed in Table 4.2 show that the model fit improves with increasing complexity, indicated by a reduction in the AIC and BIC as p increases. Note that by including climate covariates alone, the model fit explains 21% of the deviance. However, the addition of zone specific climate information to the non-climate model to form the combined model results in only an additional 6% in deviance explained. This implies that disease

Table 4.2: Deviance (D), pseudo-R^2_D, number of parameters (p), degrees of freedom ($n-p$), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for models with different subsets of covariates fit using the negative binomial GLM.

Model	Deviance	R^2_D	p	$n-p$	AIC	BIC
Null model	63007	0	2	60262	404321	404330
Climate model	61882	0.21	5	60259	389550	389586
Non-climate model	61495	0.33	99	60165	380425	381308
Combined model	60520	0.39	123	60141	374515	375614

models that use climate covariates alone may over emphasise the importance of these covariates and in fact part of the explained deviance could be captured by non-dynamical confounding factors such as altitude or area-specific annual cycle terms. The best model that emerged from the investigation was the combined model:

$$\begin{aligned} y_{st} &\sim \text{NegBin}(\mu_{st}\,, \kappa) \\ \log \mu_{st} &= \log e_{st} + \alpha + \delta_{1t'(t)} + \delta_{2k(s)} + \delta_{3k(s)t'(t)} + \sum_j \gamma_j w_{jst} \\ &\quad + \sum_j \beta_j x_{jst} + \sum_j \beta_{jk(s)} x_{jst} \end{aligned} \tag{4.5}$$

The inclusion of the factors reflecting zone, month and interaction between zone and month allow the baseline of the model to vary depending on which zone and calendar month is of interest. The interaction of zone with the three climate covariates allows the slope parameter $\beta_{jk(s)}$ to vary depending on the geographic setting. For example, to find the dengue relative risk for the microregions in the Amazon Rainforest in August 2001 with $k(s) = 1$, $t'(t) = 1$ and $t = 8$:

$$\log \rho_{s,8} = \alpha + \sum_j \gamma_j w_{j,s,8} + \sum_j \beta_j x_{j,s,8}.$$

To find the dengue relative risk for the microregions in the South Atlantic Rainforest in March 2008 with $k(s) = 7$, $t'(t) = 8$ and $t = 87$:

$$\log \rho_{s,87} = \alpha + \delta_{1,8} + \delta_{2,7} + \delta_{3,8,7} + \sum_j \gamma_j w_{j,s,87} + \sum_j \beta_j x_{j,s,87} + \sum_j \beta_{j,7} x_{j,s,87}.$$

Extra heterogeneity in the data, which is not satisfactorily explained by the month or zone factors and associated interactions, is handled by the scale parameter κ in the negative binomial. From the model fit, κ was estimated to be 0.32 with standard error

Table 4.3: Parameter estimates $\beta_{jk(s)}$ (standard error) for climate covariates in the 8 zones of Brazil (j is the parameter index and k(s) is the zone index).

Zone	precipitation ($\beta_{1k(s)}$)	temperature ($\beta_{2k(s)}$)	ONI ($\beta_{3k(s)}$)
Amazon Rainforest ($k(s) = 1$)	-0.005 (0.007)	**-0.217** (0.019)	**-0.157** (0.034)
Caatinga ($k(s) = 2$)	**-0.070** (0.009)	-0.02 (0.029)	-0.018 (0.054)
Cerrado ($k(s) = 3$)	**0.068** (0.01)	**0.135** (0.028)	**-0.408** (0.055)
North East Atlantic Rainforest ($k(s) = 4$)	**0.196** (0.02)	**0.089** (0.039)	**-0.223** (0.065)
Pampa ($k(s) = 5$)	-0.003 (0.07)	**0.347** (0.12)	**-0.357** (0.174)
Pantanal ($k(s) = 6$)	**0.437** (0.112)	**0.384** (0.126)	**-1.345** (0.187)
South East Atlantic Rainforest ($k(s) = 7$)	**0.041** (0.014)	**0.466** (0.029)	**-0.611** (0.055)
South Atlantic Rainforest ($k(s) = 8$)	**0.337** (0.019)	**0.85** (0.031)	-0.096 (0.064)

Estimates in bold face are significant at the 0.05 level.

0.002, confirming a mean variance relationship considerably different from that of the Poisson (equal mean and variance, $\kappa = \infty$). This further justifies the use of a negative binomial rather than a Poisson GLM for model selection purposes.

The estimated parameters and standard errors for the climate variables for each zone, included in the final combined model, are listed in Table 4.3. For the reference zone (Amazon Rainforest), the coefficient estimate for precipitation is $\hat{\beta}_1$ whereas for the other zones the coefficient estimate is $\hat{\beta}_1 + \hat{\beta}_{1k(s)}$ with $k(s) = 2, \ldots, 8$. The following formula was used to estimate the standard errors for the climate coefficient estimates in each zone:

$$Var[\hat{\beta}_j + \hat{\beta}_{jk}] = Var[\hat{\beta}_j] + Var[\hat{\beta}_{jk}] + 2Cov[\hat{\beta}_j, \hat{\beta}_{jk}], \tag{4.6}$$

where $Cov[\cdot]$ is the covariance of $\hat{\beta}_j$ and $\hat{\beta}_{jk}$. The standard errors were found by taking the square root of Equation 4.6. At the 0.05 level of significance, precipitation was found to have a statistically significant positive association with dengue relative risk in the South and South East Atlantic Rainforest, Pantanal and Cerrado zones. There was a statistically significant negative relationship in the Caatinga zone and no significant relationship was found in the Amazon Rainforest and Pampa zones. There was a statistically significant positive association with temperature everywhere except for the Amazon Rainforest (negative relationship) and Caatinga (no relationship).

A positive relationship between temperature/precipitation and dengue may be the result of warm and humid conditions promoting mosquito development and rain water filling

discarded containers outdoors to create mosquito breeding sites. Therefore, an epidemic could be more likely if the temperature and/or precipitation in the preceding months are above average. However, the effects are not uniform across zones. The term $\exp(\beta_j)$ is the relative risk for a unit increase in x_{jst}. From the model fit, a 1mm per day increase in average rainfall over a 3 month period would result in approximately a 4% ($\exp(0.041) = 1.04$) increase in dengue relative risk the following month in the South East Atlantic forest and a 22% increase in the North East Atlantic Rainforest. Similarly, a 1°C increase in the temperature over a 3 month period would result in a 59% increase in dengue relative risk the following month in the South East Atlantic Rainforest and a 9% increase in the North East Atlantic Rainforest. The negative relationship between temperature and dengue relative risk in the Amazon Rainforest could be attributed to cooler than average temperatures providing optimum conditions for mosquito reproduction. The negative relationship between precipitation and dengue relative risk in Caatinga could be due to the disruption of breeding sites by increased rainfall (see Chapter 2, Section 2.2).

The ONI has a negative and statistically significant association with dengue relative risk in all zones bar Caatinga and South Atlantic Rainforest. This is because the major dengue epidemics, in 2002 and 2008 in particular, were preceded by negative SST anomalies in the Niño 3.4 region. The importance of the ONI as a predictor for dengue relative risk in Brazil will be explored further in Section 4.7.

As expected from Chapter 3, altitude had a statistically significant negative association with dengue relative risk ($\gamma_1 = -0.157$, standard error = 0.000041) and population density had a statistically significant positive association ($\gamma_2 = 0.077$, standard error = 0.0187). The parameter estimates and approximate 95% confidence intervals for the month factor $\delta_{t'(t)}$ and zone factor $\delta_{k(s)}$ are shown in Figure 4.7. Note that the parameter estimate for the month August ($t'(t) = 1$) and zone Amazon Rainforest ($k(s) = 1$) is the paramter estimate for the intercept α. The interaction terms $\delta_{t'(t)k(s)}$ between month and zone are not included but were statistically significant at the 0.05 level.

Figure 4.8a shows the observed and modelled dengue incidence rate (DIR, see Chapter 3, Section 3.2.2, Eqn. 3.1) for all months and locations in Brazil. Although there is a large scatter, the local polynomial regression curve indicates an overall positive association between the observed and model fit DIR. Figure 4.8b illustrates the temporal pattern of observed and model fit DIR, for dengue counts averaged across all the microregions

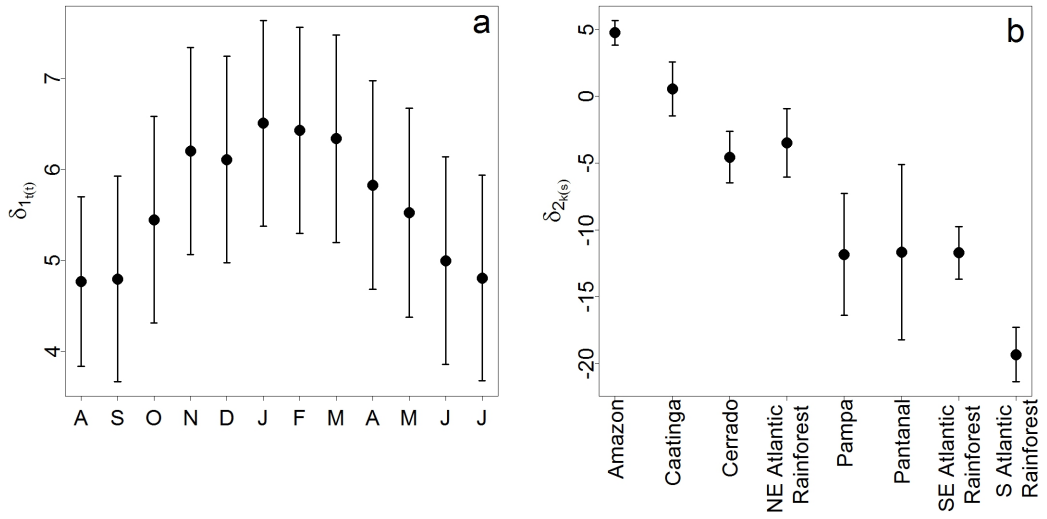


Figure 4.7: Parameter estimates (circle) and approximate 95% confidence intervals (bars) for (a) month and (b) zone factors.

in Brazil. Fitted values are displayed, including spatio-temporal climate variability and holding climate constant (e.g. precipitation, temperature and ONI set to mean values in time and space). This separates the contribution of the climate effects and the annual cycle to the estimated temporal variation in dengue. By including climate variation, an improved overall fit is achieved in the austral summer of 2005 and 2008, compared to fitted values driven by the annual cycle. However, in 2001, the inclusion of climate information caused the model to overestimate the dengue incidence rate.

Figures 4.9 and 4.10 show a break down of the comparison between the observed and modelled dengue incidence rate across the eight geographic zones. Despite the large scatter, the model appears to represent observed DIR best in the Cerrado (Fig 4.9c) and South East Atlantic Rainforest (Fig 4.9g) zones. The inclusion of climate information allowed the model to correctly estimate an increase in DIR in 2008 in Cerrado (Fig 4.10c) and South East Atlantic Rainforest (Fig 4.10g) zones, in 2007 in Caatinga zone (Fig 4.10b) and 2002 in the North East Atlantic Rainforest (Fig 4.10d). However, there are many instances where the inclusion of climate information does not help to estimate the occurrence of zone specific dengue epidemics.

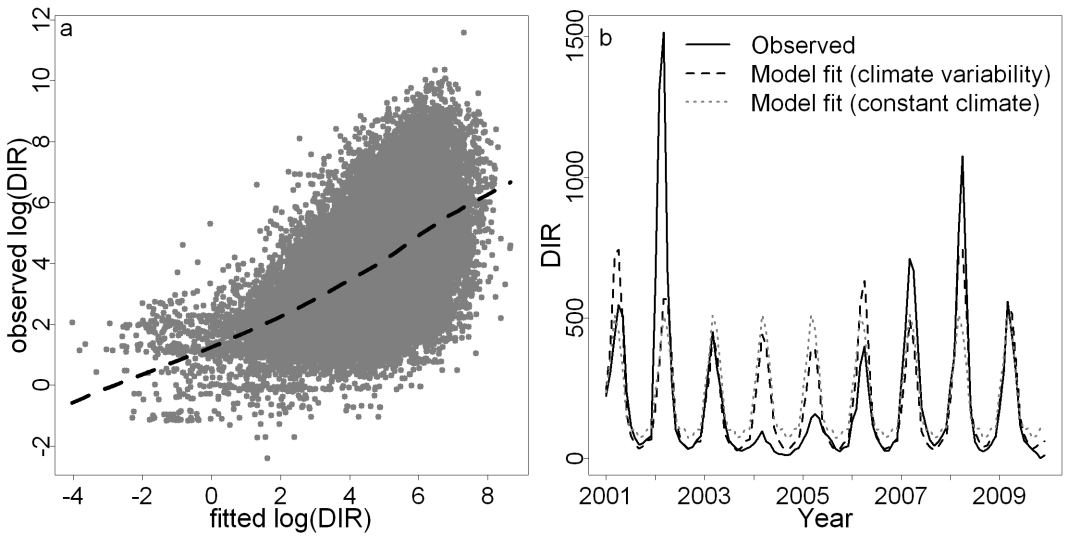


Figure 4.8: (a) Observed and model fit DIR for all months (108) and microregions (558) in Brazil. Dashed curve - local polynomial regression fit. (b) Total observed (solid line), model fit with climate variability (dashed line) and model fit holding climate constant in time and space (dotted line) dengue incidence rate (DIR) from January 2001 - December 2009.

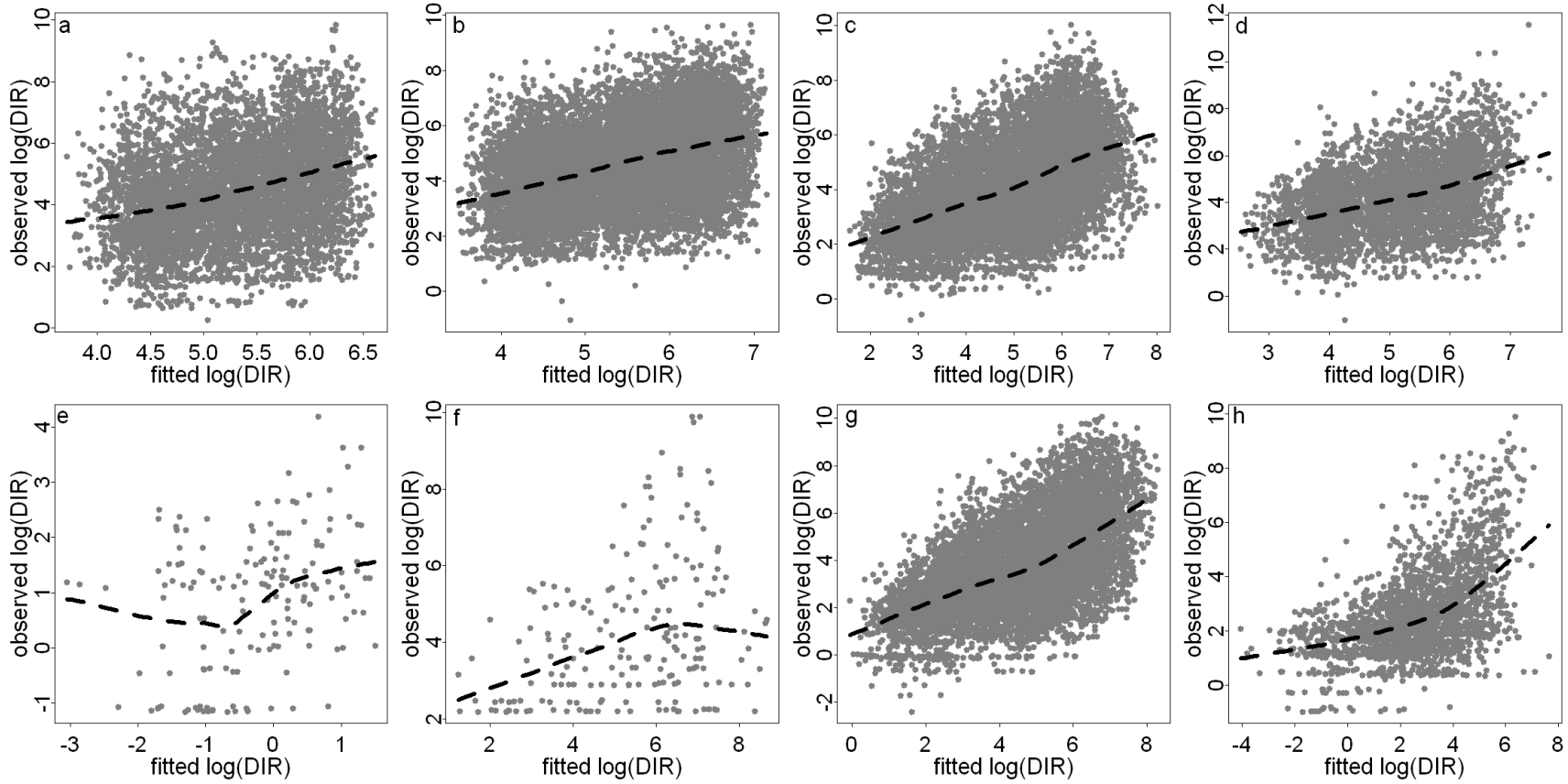


Figure 4.9: Observed and model fit DIR at the linear predictor level for all months (108) and microregions within (a) Amazon Rainforest, (b) Caatinga, (c) Cerrado, (d) North East Atlantic Rainforest, (e) Pampa, (f) Pantanal, (g) South East Atlantic Rainforest and (h) South Atlantic Rainforest. Dashed curve - local polynomial regression fit.

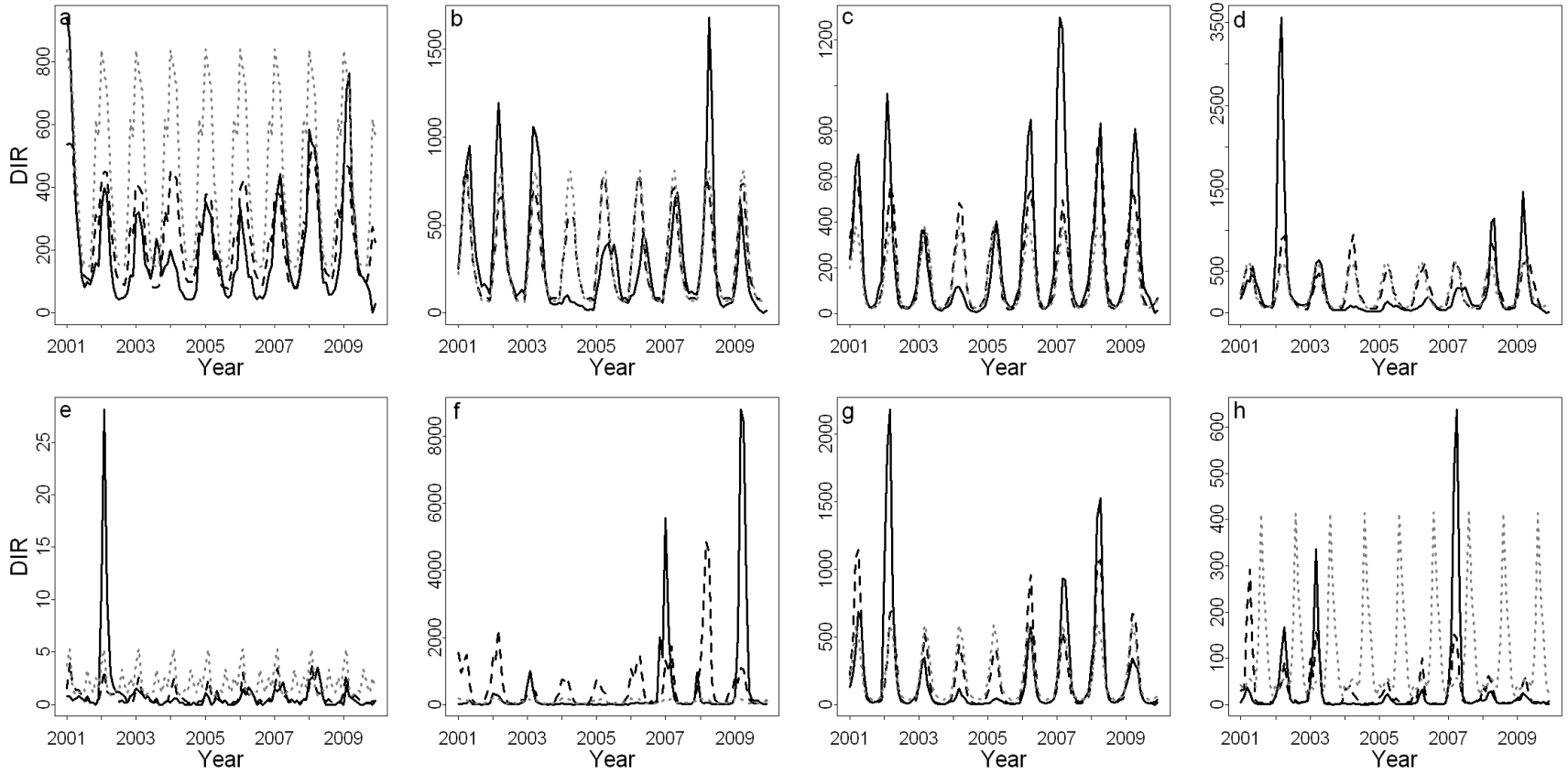
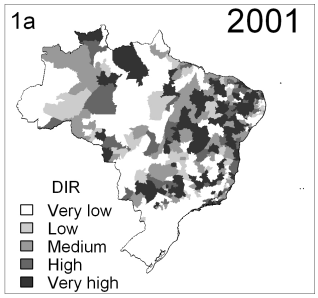
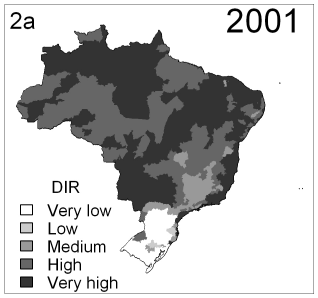


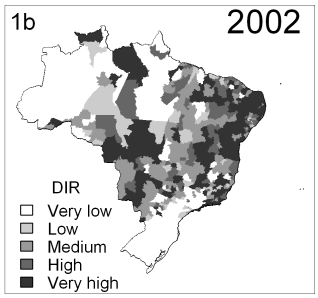
Figure 4.10: Space-averaged observed (solid line), model fit with climate variability (dashed line) and model fit holding climate constant in time and space (dotted line) dengue incidence rate (DIR) from January 2001 - December 2009 for (a) Amazon Rainforest, (b) Caatinga, (c) Cerrado, (d) North East Atlantic Rainforest, (e) Pampa, (f) Pantanal, (g) South East Atlantic Rainforest and (h) South Atlantic Rainforest.

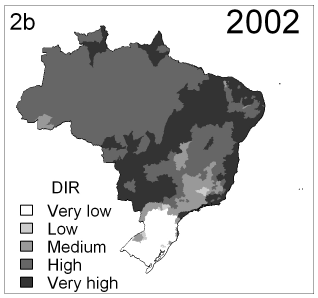
One important aspect of such a model to a public health decision maker is its ability to predict dengue during the peak dengue season which, for several zones, occurs in February-April (FMA). In Figure 4.11, the relationship between observed and model fit dengue incidence rate for the peak dengue season FMA for the 9 years 2001-2009 across all Brazilian microregions is illustrated. The model produces some systematic errors. For example, DIR is consistently over-predicted in the Amazon region each year. This may be partly due to inadequate health care provision and reporting in this sparsely populated region. However, the GLM correctly predicts low DIR in the South region. As the climate in this region is temperate, temperature and precipitation conditions are not conducive to the proliferation of the dengue mosquito. In 2001, the GLM incorrectly predicts high DIR for many areas (see Fig. 4.11.1a and 2a). Overall, the GLM is able to correctly capture very high DIR in some locations in 2002 (see Fig. 4.11.1b and 2b) and 2008 (see Fig. 4.11.1h and 2h), when serious epidemics occurred. In, 2003-2005, the model is also able to correctly predict lower DIR than for other years, particularly in the South East region (see Fig. 4.11.1c, 2c, 1d, 2d, 1e, 2e).

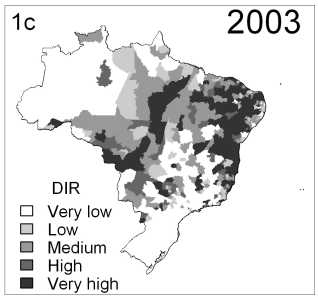
The influence of the climate variables for this season is demonstrated in Figure 4.12a, which shows the time series of observed DIR for the FMA season, model fit DIR, holding climate information constant (dotted line) and allowing climate to vary (dashed line). Note that the solid (observed) and dashed (model fit with climate variability) lines summarise the information presented in the maps in Figure 4.11.1 and Figure 4.11.2, respectively. As climate variables are the only source of temporal information in the model, omitting them results in the same prediction for every month/season of each year (dotted line, Fig. 4.12a). Climate information allows some of the temporal variability to be captured, for example between 2007-2009 (dashed line, Fig. 4.12a). Figure 4.12b and c compare the spatial distribution of observed and modelled DIR for the FMA season in 2008, a year when the model correctly predicts an overall increase in DIR. The GLM is unable to fully capture the spatial variability in DIR, particularly in the Amazon region and the North East. There is some agreement between observed and modelled DIR in the South East region and the model partially captures very low incidence rates in South Brazil.

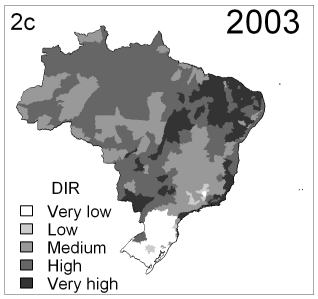


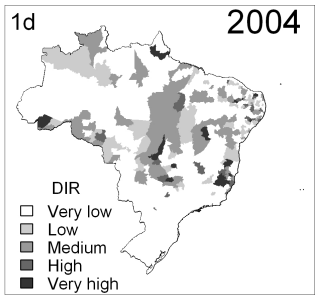


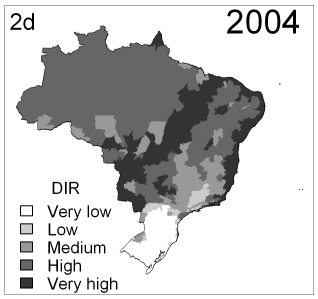


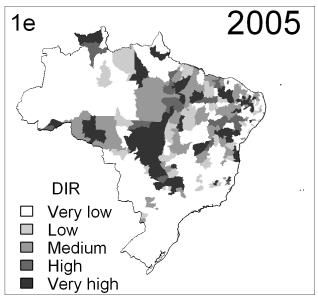


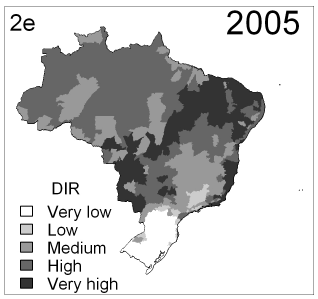












continued overleaf

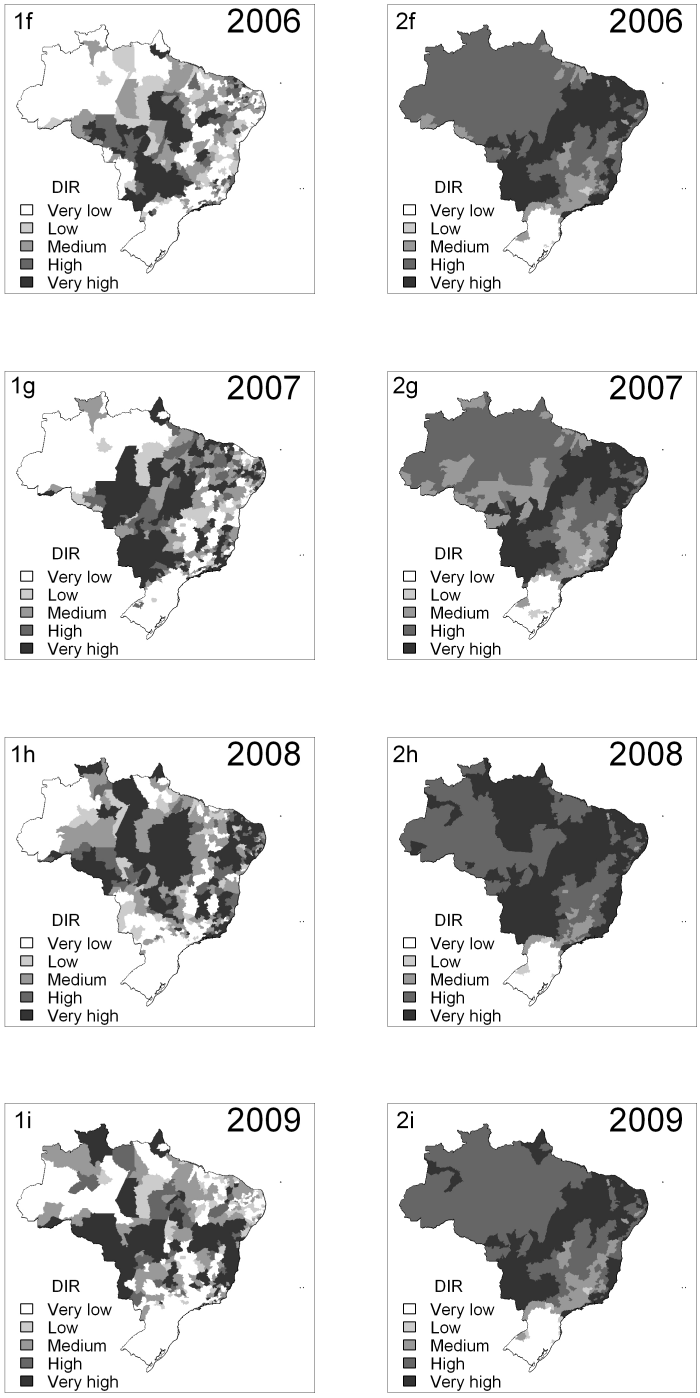


Figure 4.11: Observed (column 1) and model fit (column 2) DIR for FMA in (a) 2001, (b) 2002, (c) 2003, (d) 2004, (e) 2005, (f) 2006, (g) 2007, (h) 2008 and (i) 2009. Category boundaries defined by 50, 100, 300 and 500 cases per 100,000 inhabitants.

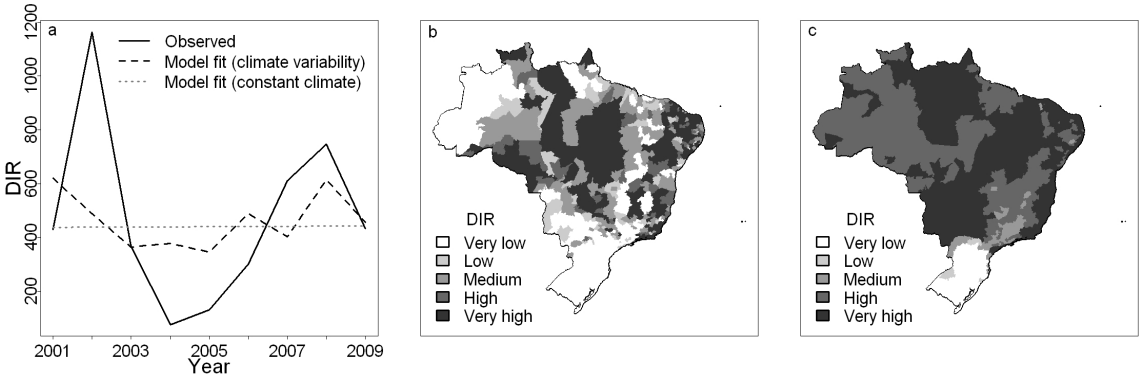


Figure 4.12: (a) Observed (solid line), model fit with climate variability (dashed line) and model fit with climate held constant (dotted line) DIR, FMA 2001-2009, Brazil. (b) Observed and (c) model fit DIR for the microregions of Brazil, FMA season 2008. Categories defined by 50, 100, 300 and 500 cases per 100,000 inhabitants.

In Figure 4.13b and Figure 4.13c the observed and fitted FMA DIR is extracted for the South East region (zones: Cerrado and South East Atlantic Rainforest) and North East Region (zones: Amazon Rainforest, Caatinga, Cerrado, North East Atlantic Rainforest). Despite the large scatter, there is a general positive association between observed and model fit dengue incidence rates in the South East (Fig. 4.13b), however the model fails to reflect observations in the North East Region (Fig 4.13c).

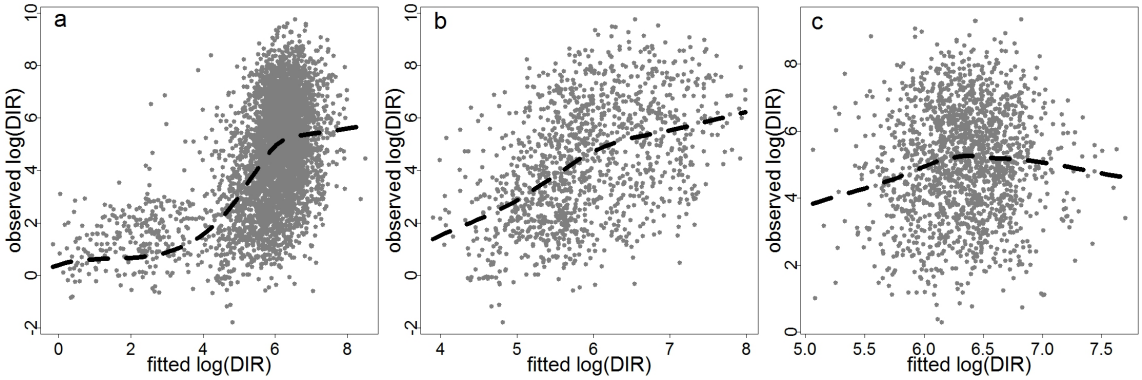


Figure 4.13: Observed and model fit DIR for 3-month season FMA 2001-2009 for (a) Brazil, (b) South East region and (c) North East region. Dashed curve - local polynomial regression fit.

Figure 4.14 shows the temporal skill of the model for the FMA season 2001-2009 in the South East and North East regions. In the South East (Fig 4.14a), the model correctly estimated a decrease in the DIR for FMA season from 2002-2003 and the peak between 2007-2009 (i.e. the 2008 epidemic). However, the model estimated a decrease in the dengue incidence rate between 2001-2002 and 2006-2007 when an increase was actually observed. In the North East (Fig 4.14b) there is a slight peak in the modelled dengue incidence rate for the 2002 and 2008 epidemics. However, between 2003-2005 the model estimated an increase in DIR when a decrease was actually observed.

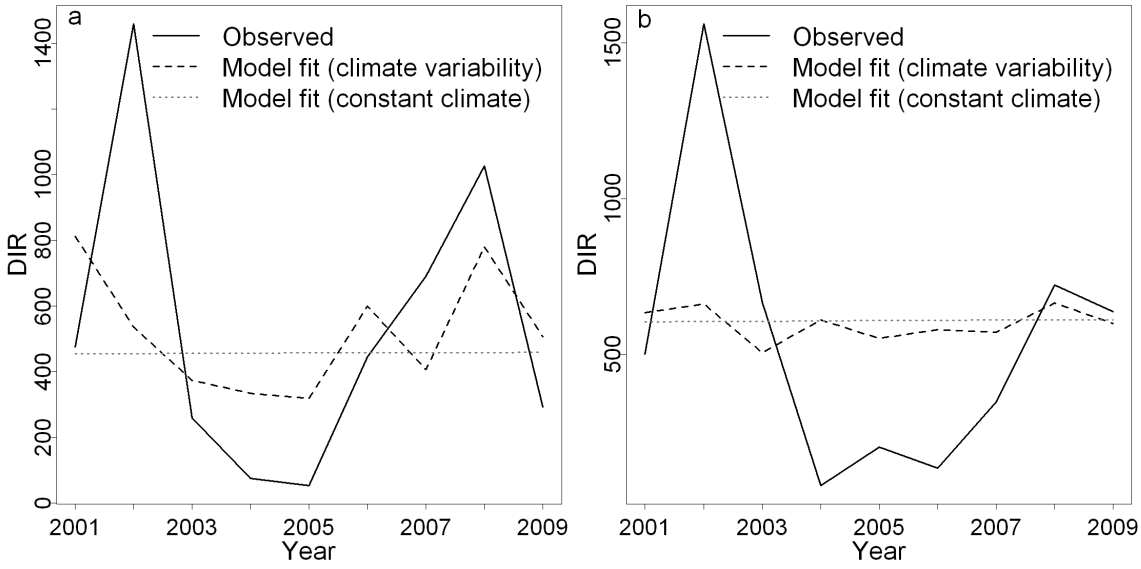


Figure 4.14: Total observed (solid line), model fit with climate variability (dashed line) and model fit with climate held constant (dotted line) DIR for FMA 2001-2009 for (a) South East and (b) North East Brazil.

Dengue early warnings would be of most use to decision makers at the microregion level, to allow effective interventions such as targeted mosquito control programmes and the allocation of resources to hospitals in microregions expecting overwhelming volumes of patients. Figure 4.15 zooms in at the microregion level to Rio de Janeiro, located in the South East region (Fig 4.15a) and Salvador da Bahia in the North East region (Fig 4.15b). A successful epidemic warning would have been possible for the FMA 2008 in Rio de Janeiro and in FMA 2002 in Salvador da Bahia. However, potential false alarms could have been issued in 2006 in Rio de Janeiro and 2004 in Salvador da Bahia, due to the inclusion of climate in the model.

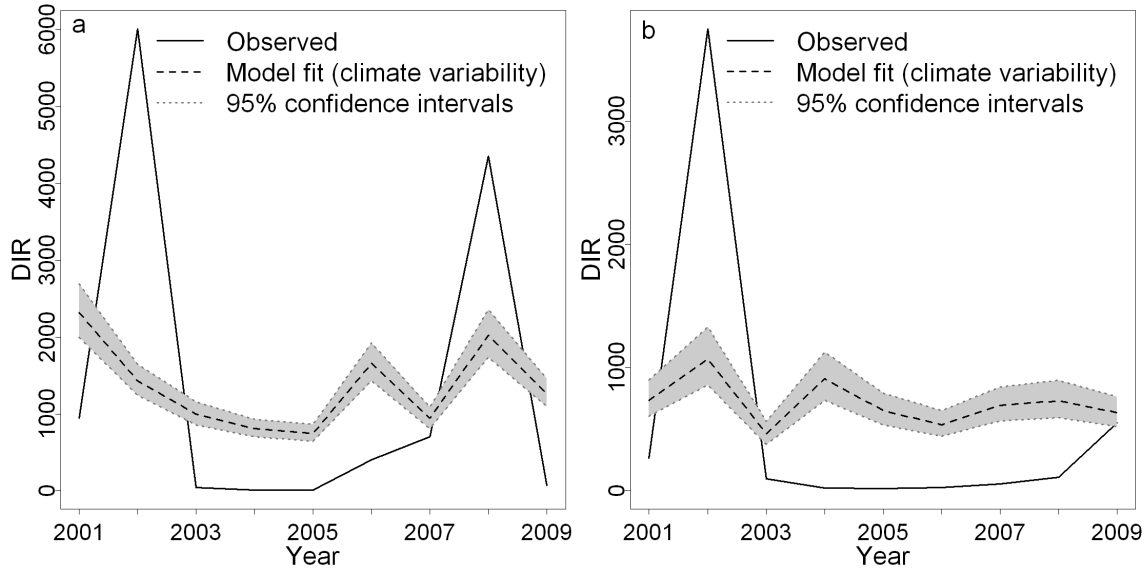


Figure 4.15: Time series of total observed DIR (solid line) and model fit DIR with climate variability (dashed line) with 95% confidence intervals (dotted line) for FMA 2001-2009 for (a) Rio de Janeiro and (b) Salvador da Bahia.

Results thus far have indicated that the model performs best in the South East region. In the North East region, a variety of complex relationships exist. This region is composed of four zones (Amazon Rainforest, Caatinga, Cerrado, North East Atlantic Rainforest). The relationship between precipitation and temperature and dengue relative risk differs between these zones. Also, exploratory analyses in Chapter 3 indicated that ENSO affects precipitation in contrasting ways across the North East region. Analyses in subsequent chapters will focus on South East Brazil; a region where dengue is most prevalent and there are a large number of densely populated urban centres, which could benefit from a climate informed dengue early warning system. This is also the region where the GLM captured some of the observed spatial variability in dengue incidence rates (see Figure 4.12c). Before extending the modelling framework in the following chapter, the relationship between ENSO and dengue incidence in the South East region, will be investigated in more detail.

4.7 Robustness of ENSO effect on dengue

During the study period, the ONI strongly influenced temporal variation in modelled dengue in the South East region and helped capture the dengue epidemic in 2008. However, the extent to which ENSO could be a driving force behind dengue epidemics remains unclear. Results from the previous chapter indicated a positive association between lagged ONI and precipitation/temperature in South East Brazil (see Fig 3.16). As precipitation and temperature are positively associated with DIR in this region, it is curious to then see a negative association between ONI and DIR. There are several possible explanations for this unexpected relationship. Firstly, the time lag between ONI and the local response in temperature and precipitation may differ between regions and for the GLM fitted to Brazil as a whole, the most suitable time lag chosen in the model selection process may not be applicable to individual regions. Secondly, the statistical association may be distorted by influential values or leverage and may not be robust between all values. Alternatively, the statistical relationship could be spurious, for example, due to the coincidence of the 2007-8 La Niña and the 2008 epidemic in Brazil (e.g. see Johansson et al., 2009a), which may actually have been caused by unobserved confounding effects such as the circulation of a new serotype. The inclusion of all months versus peak months in the model framework will also be investigated.

4.7.1 Regional model framework

To investigate the robustness of the ONI as a predictor for dengue, a simplified GLM for the South East region of Brazil with only ONI and the annual cycle as exploratory variables was formulated. Initial experiments revealed that there was little difference in the ONI coefficient estimate between the two zones in South East Brazil (south portion of the Cerrado zone and South East Atlantic Rainforest). Therefore, zone was not included as a categorical variable in the model. The model specification is as follows:

$$\begin{aligned} y_{st} &\sim \text{NegBin}(\mu_{st}\,, \kappa) \\ \log \mu_{st} &= \log e_{st} + \alpha + \delta_{t'(t)} + \beta x_t, \end{aligned} \tag{4.7}$$

where x_t is the ONI. This model was used to find the most suitable time lag for the ONI in the South East region, according to the time lag which gave the lowest AIC. Using this time lag, the model was assessed for outliers and leverage points. Potentially

influential observations were removed (as described below) and the model was refitted to see if certain points had a large effect on the outcome of the parameter estimates.

4.7.2 Local variations in optimal time lag

For the model selection process for the GLM fitted to the whole of Brazil (see Section 4.6), stepwise regression favoured ONI lagged 6 months before dengue incidence (4 months behind precipitation and temperature). However, ENSO may affect different regions of Brazil at different time lags. The simplified GLM in Equation 4.7 was fitted repeatedly by replacing x_t with the ONI at time lags of 2 to 12 months previous to dengue (0-10 months previous to precipitation/temperature). All estimates were significant at the 0.05 level using a z-test. For the South East region, as the lag increased to 6 months, the AIC decreased (see Table 4.4 and Figure 4.16). This is consistent with findings from the combined model (Eqn. 4.5), fitted to the whole of Brazil. However, beyond 6 months, the AIC waxes and wanes between 7 and 12 months lag with a minimum at 8 months lag and a subsequent peak at 10 months lag. ONI lagged 6 months previous to dengue was preferred as the AIC consistently decreases to this lag and process understanding between ENSO, local climate in South East Brazil and dengue suggests that this time lag is sensible. For example, when considering the peak dengue season for South East Brazil (FMA), the August-October (ASO) ONI would be a good indicator of ENSO phase for the forthcoming seasons, as El Niño or La Niña events in the Pacific Ocean are typically established by August. Prior to this, it can be difficult to determine how ENSO will evolve. This drop-off in monthly persistence (lagged correlation) is often referred to as the spring predictability barrier (see Torrence and Webster, 1998 for further details).

The total DIR for the FMA season in the South East region, extracted from the 'Brazil' model, compared to the DIR from the simplistic 'South East' model are similar (see Fig. 4.17). This suggests that subsequent inference may be relevant to a more complex model fitted to the South East region of Brazil (see Chapter 5).

Table 4.4: Summary of parameter estimates for ONI and AIC at different time lags between dengue relative risk and ONI for South East Brazil.

Time lag	ONI coefficient estimate	AIC
2	-0.392	104322
3	-0.4176	104293
4	-0.456	104246
5	-0.497	104192
6	-0.522	104158
7	-0.538	104142
8	-0.548	104138
9	-0.535	104157
10	-0.528	104163
11	-0.527	104154
12	-0.531	104139

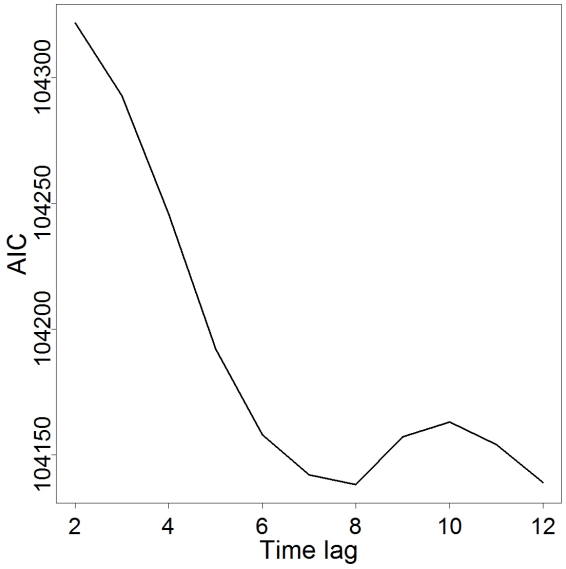


Figure 4.16: Change in AIC with increasing time lag from 2 to 12 months between dengue relative risk and ONI for South East Brazil.

4.7.3 Influence and leverage

To investigate if the reported association between dengue relative risk and ENSO is robust or if one ENSO event, for example, had a large effect on the outcome of parameter estimates, the model fit was examined for influential observations and leverage points.

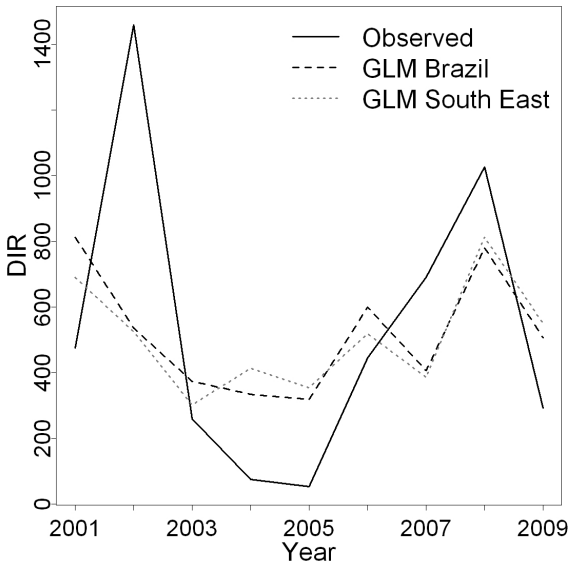


Figure 4.17: Total observed (solid line), GLM fit, from Brazil model (dashed line) and GLM fit, from South East simplified (ONI and annual cycle) model (dotted line) DIR for FMA 2001-2009.

If removing an observation from the analysis results in a substantial modification of the parameter estimates, the observation is said to be influential. For a linear model, a projection matrix known as the hat matrix can be computed

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}', \tag{4.8}$$

where $\mathbf{X}$ is the matrix of explanatory variables, known as the design matrix. $\mathbf{H}$ relates fitted values $\hat{\boldsymbol{\mu}}$ to observed values $\mathbf{y}$ via

$$\hat{\boldsymbol{\mu}} = \mathbf{H}\mathbf{y}.$$

$\mathbf{H}$ describes the influence each observed value has on each fitted value (Hoaglin and Welsch, 1978). The diagonal elements of $\mathbf{H}$ are often called the leverages (Draper and Smith, 1998) since h_{ii} indicates how heavily y_i contributes to μ_i.

Leverages are slightly different for GLMs. The IRLS algorithm used to fit GLMs uses weights w_i, which are a function of the fitted values $\hat{\boldsymbol{\mu}}$ and inversely proportional to the variance function $V(\hat{\mu}_i)$ (see Appendix A). By forming a matrix $\mathbf{W} = diag\{w_i\}$ the hat matrix becomes

$$\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{\frac{1}{2}},$$

equivalent to replacing $\mathbf{X}$ by $\mathbf{W}^{\frac{1}{2}}\mathbf{X}$ in the linear model version in Equation 4.8 (McCullagh and Nelder, 1989). One important difference from the linear model case is that the

leverages are no longer just a function of $\mathbf{X}$ and now depend on the response through the weights $\mathbf{W}$ (Faraway, 2006). This effectively allows for the change in variance with the mean. The leverages combined with Studentised deviance residuals, given by

$$r_{(SD)i} = \frac{r_{(D)i}}{\sqrt{\hat{\varphi}(1-h_{ii})}},$$

provides a means of identifying exceptional data points (Hoaglin and Welsch, 1978). An observation which has both high influence and a large residual is influential and may distort the accuracy of the model fit and predictions. In practice, observations are typically considered high leverage if $h_{ii} > 3(p+1)/n$ (Krzanowski, 1998).

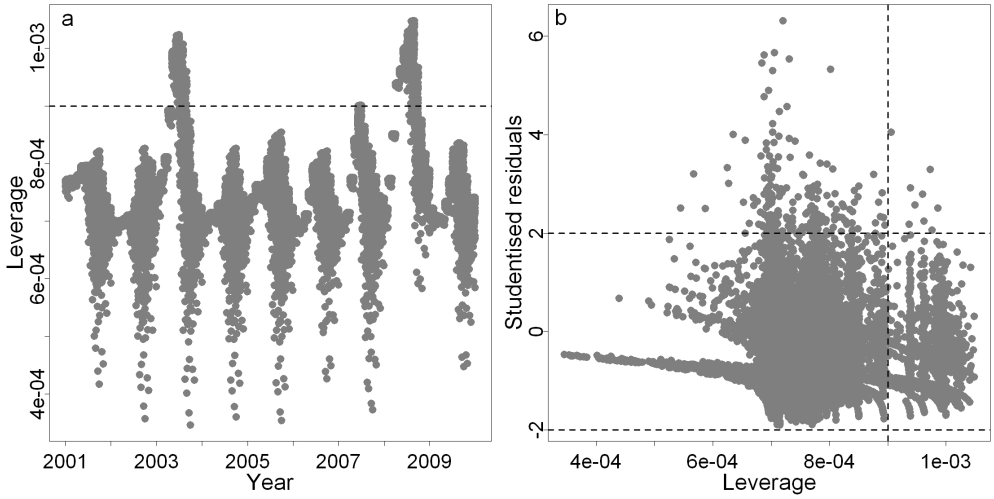


Figure 4.18: (a) Leverage over time and (b) Studentised residuals against leverage for South East GLM. Dashed lines $h_{ii} = 0.0009$ and $\left|r_{(SD)i}\right| > 2$: threshold for removal of points.

For the South East GLM, this threshold was found to be 0.0026, whereas the maximum $h_{ii} = 0.001$. Therefore, the model fit is not considered to be affected by influential observations. However, Figure 4.18a shows that data points in 2003 (preceded by an El Niño event) and 2008 (preceded by /coincident with a La Niña event) exhibited relatively greater leverage than others. The plot of Studentised residuals versus leverage (Fig. 4.18b) has a vertical line that indicates relatively high leverage points and two horizontal lines that indicate potential outliers. Potential outliers ($\left|r_{(SD)i}\right| > 2$) and relatively high leverage points ($h_{ii} > 0.0009$) identified from this plot were removed from the dataset and the model was refitted. The coefficient estimate for ONI did not change much in value and remained statistically significant at the 0.05 level (see Table 4.5). This suggests that ONI is a statistically robust predictor for dengue relative risk.

4.7.4 Peak months

Another potential issue is the inclusion of all calendar months in the analysis. For epidemic prediction, it is desirable to include all months in the analysis, to be able to detect the potentially early onset of an epidemic and to be able to capture rare unexpected/out of season events in future years. However, the relationship between ENSO and climate across Brazil is not consistent throughout the year and the inclusion of all months may distort the relationship between ENSO and dengue. Accordingly, the months June-December where removed from the dataset, leaving the months January-May, when dengue relative risk peaks (see Fig 3.8). The model was refitted to the peak months (January - May) only. Again, the ONI coefficient remained negative and highly statistically significant and its value did not dramatically change (see Table 4.5).

Table 4.5: Coefficient estimates for ONI by deletion of points for South East Brazil.

Data	ONI coefficient estimate	Standard error	P-value
All months	-0.522	0.028	2×10^{-16}
Exclude leverage/outliers	-0.506	0.029	2×10^{-16}
Peak months	-0.487	0.042	2×10^{-16}

4.7.5 Summary

The results from this section suggest that the association between ONI and dengue relative risk found in the GLM for Brazil is not statistically spurious and that the analysis was not affected by leverage, outliers or the inclusion of all months of the year. The possibility still remains that ENSO events coincidently occurred at the same time as some other unobserved confounding factor in the disease system. However, it is clear that a unique relationship exists between anomalous ENSO events and dengue relative risk during the time period. To ignore this evidence would be misleading. Therefore, the index will remain in the model framework in the following chapter, where allowances will be made for unobserved confounding factors via the inclusion of random effects in the linear predictor. These may account for factors such as serotype introduction or public health interventions and could potentially render the ONI an insignificant predictor for dengue risk.

4.8 Conclusion

In this chapter, a negative binomial GLM was identified as a suitable model for dengue counts in Brazil, using a range of statistical criteria. The data were found to display substantial overdispersion compared to a Poisson model. Alternative approaches could have been adopted to account for overdispersion. For example, by moving away from the complete distributional specification of the Poisson model to the specification of the quasi-Poisson model (Wedderburn, 1974). This approach allows the dispersion parameter φ, to be greater than one to inflate the standard estimates of the covariate estimates. However, a negative binomial distribution, that permits more flexible modelling of the variance than the Poisson, was instead specified. This allowed formal statistical tests to be conducted for model comparison. The negative binomial model successfully accounted for the unexplained overdispersion in the Poisson model.

The best model that emerged from the investigation comprised a combination of climate and non-climate covariates and a selection of interactions to explain dengue relative risk in Brazil. The selected climate variables were 3-month averaged precipitation and temperature (lagged two months) and the ONI (lagged 6 months). The residual deviance when using climate covariates alone ($R^2_D = 21\%$) was considerably greater than the additional contribution of climate to the residual deviance when combined with non-dynamical confounding factors (non-climate model: $R^2_D = 33\%$, combined model $R^2_D = 39\%$). This finding is important for the assessment of a climate-based epidemic early warning system as the incorporation of dynamic climate information is costly, thus a fair assessment of the contribution of climate variability to the prediction of disease risk is essential.

The inclusion of the ONI accounts for some of the inter-annual variation in dengue incidence rates, particularly during the 2008 dengue epidemic. However, the relationship between ONI and DIR is not consistent with that expected from process understanding of the effects of temperature and precipitation on dengue. The surprising relationship between ONI and dengue relative risk in the South East region of Brazil was further investigated and ONI appeared to be a statistically robust predictor for dengue. However, ENSO events may have coincidently occurred with some other factor. For example, inter-annual variability in DIR may be attributable to factors such as population immunity to the dominant circulating serotype or specific health interventions and vector control

measures. Unfortunately, information regarding these aspects of the disease system is not readily available. The model also poorly reproduces spatial variability across microregions. Therefore, the use of random effects may be valuable to allow for unobserved latent structures in the model (McCulloch and Searle, 2001). For example, to capture the impact of unknown/unobserved confounding factors, such as the introduction of a new dengue serotype in a certain area of Brazil, or to account for misreporting of dengue cases.

An important characteristic of GLMs is that they assume independent (or at least uncorrelated) observations, but this assumption may not be valid. There could be strong temporal correlation effects within some areas and there could also be spatial clustering effects in neighbouring microregions. To allow for such latent effects and correlation structures, the fixed effects GLM is used as a starting model and further refined in the next chapter by including random effects in the modelling framework. In the following chapter, a generalised linear mixed model (GLMM) will be adopted and implemented in a Bayesian framework using Markov Chain Monte Carlo (MCMC), to better capture spatio-temporal variations in dengue risk. It is hoped that the use of a mixed effect model will help to gain an understanding of the sensitivity of dengue risk to climate variability in South East Brazil.

Chapter 5

Extension to a Bayesian hierarchical model framework

5.1 Introduction

The aim of this chapter is to extend the modelling framework developed in the previous chapter to a generalized linear mixed model (GLMM). GLMMs (Breslow and Clayton, 1993) are extensions of GLMs that allow for additional variation in the response arising from unobservable random effects. In Chapter 4, a negative binomial GLM (fixed effects) was tested which allowed for extra-Poisson variation, or overdispersion, in the count data via the additional parameter κ. Although the GLM attempted to account for confounding factors by the inclusion of non-climate variables and interactions between the annual cycle, geographic zones and climate covariates, there was still a large amount of unexplained variation. Before inference can be made as to the sensitivity of dengue risk to climate variability in South East Brazil, GLMMs including unstructured and structured random effects in the linear predictor (mixed effects) will be explored. Such models will help to ascertain how much dengue variation can be attributed to climate, as previously, variation in dengue may have been incorrectly attributed to climate rather than unobserved spatio-temporal factors associated with dengue risk. The inclusion of random effects introduces an extra source of variability (a latent effect) into the model to capture the impact of such unknown/unobserved confounding factors. For example, the emergence of a new dengue serotype, which could vary spatially and temporally. Spatially

unstructured random effects can assist in modelling overdispersion, previously allowed for solely via the single scale parameter in the negative binomial GLM, while spatially structured random effects allow for correlated heterogeneity between microregions.

Parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, in a GLMM can be estimated in various ways, including classical likelihood-based methods. A more natural way to handle random effects is to use a Bayesian framework. This is the approach that will be adopted in this thesis. In contrast to classical inference, the Bayesian approach accounts for parameter uncertainty by assigning prior distributions to the parameters. A more detailed account of the Bayesian framework and MCMC methods can be found in Appendix B.

Hierarchical models can be created by parameterising prior distributions with unknown 'hyperparameters' $\boldsymbol{\vartheta}$ which have their own 'hyperprior' distribution $p(\boldsymbol{\vartheta})$ (Appendix B.1). Bayesian hierarchical models play an important role in modelling the complexity of data structures in spatial epidemiology (Lawson et al., 2003). These models can also take into account a spatial pattern in disease, for example, the tendency of areas within close geographical proximity to have similar disease rates. MCMC methods (Appendix B.2) make estimation of parameters in Bayesian models a practical feasibility (Gilks et al., 1996; Brooks, 1998; Gelman et al., 2004). One further advantage of the Bayesian approach is that the associated MCMC sampling yields samples from full posterior predictive distributions $p(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\mathbf{y})$ which automatically incorporate all components of variance at the different levels in the model. Therefore, a full assessment of prediction uncertainty can be more easily obtained with Bayesian MCMC estimation than with the more traditional maximum likelihood approach.

The investigation now considers modelling dengue at the region rather than country level. The South East region of Brazil was the region where the GLM performed best, dengue is most prevalent and there are a large number of densely populated urban centres, which could benefit from a climate informed dengue early warning system. Therefore, subsequent modelling will focus on the South East region.

5.2 Generalised linear mixed model framework

In this chapter, a negative binomial GLMM is developed to model dengue cases in the South East region of Brazil. The model now includes random effects in the linear predictor. The GLMM is formulated as a Bayesian hierarchical model where the first level is the negative binomial model for the dengue counts, from January 2001-December 2009 in the 160 microregions in South East Brazil. The second level models additional microregion specific variation by including random effects in the linear predictor of dengue relative risk.

The negative binomial GLMM framework is:

$$\begin{aligned} y_{st}|\mu_{st} &\sim \text{NegBin}(\mu_{st}, \kappa) \\ \log \mu_{st} &= \log e_{st} + \log \rho_{st} \\ \log \rho_{st} &= \Psi_{st} + \Lambda_{st}, \end{aligned}$$

where y_{st} is the dengue count for microregion $s = 1, \ldots, 160$ and time $t = 1, \ldots, 108$, μ_{st} is the corresponding mean dengue count and κ is the scale parameter. Again, the expected cases e_{st} are treated as an offset in the model (see Eqn. 3.2, Section 3.2.3, Chapter 3). The relative risk ρ_{st} is now composed of 'fixed effects' Ψ_{st} and 'random effects' Λ_{st} (hence mixed effects model).

5.3 Fixed effects

The fixed effects consist of

$$\Psi_{st} = \alpha + \delta_{t'(t)} + \sum_j \gamma_j w_{jst} + \sum_j \beta_j x_{jst},$$

i.e. the selected combined model for the dengue relative risk (see Chapter 4, Eqn. 4.5) with the difference that only the factor 'calendar month' $\delta_{t'(t)}$ (i.e. annual cycle) is included as a categorical variable. This model is now specific to the region (rather than country) level. There was little difference in the climate coefficient estimates between the two zones in South East Brazil (south portion of the Cerrado zone and South East Atlantic Rainforest). Therefore, for this region, the zone factor is no longer necessary. Spatial heterogeneity in the data will now be accounted for at a finer geographic resolution

via the inclusion of spatial random effects for each microregion. Again, as mentioned in Chapter 4, there was no strong evidence of non-linear relationships in the climate covariates for this particular region.

5.4 Random effects

The inclusion of random effects introduces an extra source of variability (a latent effect) into the model to capture the impact of unknown/unobserved confounding factors. The random effects Λ_{st} can include structured or unstructured, spatial, temporal, or spatio-temporal random effects or any combination of these, depending on the problem that is being modelled. The following gives a brief description of the prior distributions for individual random effects that will be considered in this chapter.

5.4.1 Spatially unstructured random effects

The simplest prior for a spatially unstructured random effect $(\Phi_1 \ldots, \Phi_S)$ assumes exchangeable random effects (unchanged by permutation of areas), given the hyperparameters $\boldsymbol{\vartheta}$:

$$p(\Phi_1, \ldots, \Phi_S|\boldsymbol{\vartheta}) = \prod_{s=1}^{S} p(\Phi_s|\boldsymbol{\vartheta})$$

where the prior distribution $p(\Phi_s|\boldsymbol{\vartheta})$ has the same form for each area s (Mollie, 1996). A typical choice for a spatially unstructured prior is a Gaussian distribution with zero mean and large variance (Clayton and Kaldor, 1987). A diffuse gamma hyperprior is often assigned to the precision (hyperparameter), i.e, the reciprocal of the variance, $\tau_\Phi = 1/\sigma_\Phi^2$. The gamma distribution has mean ζ/η and variance ζ/η^2, where ζ is the shape parameter and η is the inverse scale parameter. Therefore, the distribution for a spatially unstructured random effect Φ_s could take the form

$$\begin{aligned} \Phi_s &\sim \mathrm{N}(0, \sigma_\Phi^2) \\ \tau_\Phi &\sim \mathrm{Ga}(\zeta, \eta). \end{aligned} \tag{5.1}$$

However, this does not allow for explicit spatial dependence between y_{st}, arising through similar dengue relative risks, for example, in neighbouring densely populated urban areas as opposed to sparsely populated rural areas.

5.4.2 Spatially structured random effects

Latent spatial dependence can be incorporated in a model by including a spatially structured random effect Υ_s, instead of the spatially unstructured random effect Φ_s. A spatial dependency structure can be imposed by assuming a prior distribution for the spatial effects which takes the neighbourhood structure of the area under consideration into account. Prior information which allows for local geographical dependence causes the relative risks in an area to be shrunk towards a local mean, according to the relative risks in neighbouring areas (Mollie, 1996). A typical choice for a spatially structured prior is a conditional intrinsic Gaussian autoregressive model (CAR) (see Besag et al., 1995);

$$\Upsilon_s|\Upsilon_{r\neq s} \sim \mathrm{N}\left(\frac{\sum_{r\neq s} a_{sr}\Upsilon_r}{\sum_{r\neq s} a_{sr}}, \frac{\sigma^2_\Upsilon}{\sum_{r\neq s} a_{sr}}\right), \tag{5.2}$$

where a_{sr} are *adjacency weights* for the microregions. Here, these are taken to be simple binary values: $a_{sr} = 1$ if microregion s has a common boundary with microregion r, $a_{sr} = 0$ otherwise. For convenience, Equation 5.2 will subsequently be notated as

$$\begin{aligned} \Upsilon_s &\sim \mathrm{CAR}(\sigma^2_\Upsilon) \\ \tau_\Upsilon &\sim \mathrm{Ga}(\zeta, \eta). \end{aligned} \tag{5.3}$$

The hyperparameter $\tau_\Upsilon = 1/\sigma^2_\Upsilon$ controls the strength of the local spatial dependence. As this formulation of the CAR is improper, a 'sum to zero' constraint is applied to Υ_s ($\sum_s \Upsilon_s = 0$) and it is then advisable to take a uniform flat prior for the model intercept $\alpha \sim \mathrm{U}(-\infty, +\infty)$ (see Best et al., 1999 for more details).

5.4.3 Combination of unstructured and structured random effects

In practice, it is often unclear how to choose between a model formulation involving either spatially unstructured random effects Φ_s or spatially structured random effects Υ_s (Mollie, 1996). An intermediate distribution on the log relative risk ($\log \rho_{st}$), that ranges from prior independence to prior local dependence, has been proposed (Besag, 1993; Besag and Mollié, 1989; Besag et al., 1991). Area-specific random effects that are divided into spatially unstructured and structured components are often termed the 'convolution prior' (Best et al., 1999; Besag et al., 1991; Bailey et al., 2005). In this prior

model, a spatial random effect Φ_s is the sum of two independent components

$$\begin{aligned} \Lambda_s &= \phi_s + \upsilon_s \\ \phi_s &\sim \mathrm{N}(0, \sigma_\phi^2) \\ \upsilon_s &\sim \mathrm{CAR}(\sigma_\upsilon^2) \\ \tau_\phi &\sim \mathrm{Ga}(\zeta, \eta) \\ \tau_\upsilon &\sim \mathrm{Ga}(\zeta, \eta) \end{aligned} \tag{5.4}$$

where ϕ_s is an independent normal variable with zero mean and variance σ_ϕ^2, describing unstructured heterogeneity in the relative risks. Parameter υ_s is modelled as an intrinsic Gaussian autoregression (CAR, Eqn. 5.2), while $\tau_\phi = 1/\sigma_\phi^2$ and $\tau_\upsilon = 1/\sigma_\upsilon^2$ are hyperparameters contained in $\boldsymbol{\vartheta}$.

5.4.4 Temporally autocorrelated random effects

To account for unobserved confounding factors in time, a temporal random effect can also be included in the model framework for every time unit (e.g. Mabaso et al., 2006). Such temporal random effects are often interacted with spatial location (Knorr-Held, 2000) in order to capture unmeasured effects, e.g. the emergence of a new serotype to the region in a particular month. However, such effects would be of little use for prediction, as the temporal randomness for future months is unknown and cannot be estimated. There is reason to believe that dengue incidence exhibits temporal serial correlation. For example, dengue relative risk in March may depend on the risk in February. Accordingly, a first order autoregressive month effect $\omega_{t'(t)}$ will be tested in the model framework, where t' is calendar month. The simplest such prior is the random walk or first difference prior (Gilks et al., 1996) in which each effect is derived from the immediately preceding effect:

$$\begin{aligned} \omega_{t'(t)} &= 0 \quad \text{if} \quad t'(t) = 1 \\ \omega_{t'(t)} &\sim \mathrm{N}(\omega_{t'(t)-1}, \sigma_\omega^2) \quad \text{if} \quad t'(t) = 2, \ldots, 12 \\ \tau_\omega &\sim \mathrm{Ga}(\zeta, \eta). \end{aligned} \tag{5.5}$$

As with the spatial random effects, the hyperprior $\tau_\omega = 1/\sigma_\omega^2$ is assigned a diffuse gamma distribution.

5.5 Selection of random effects

A series of candidate negative binomial models of increasing complexity, from a fixed effects model (GLM) to a spatio-temporal mixed effects model (GLMM) were tested. The model parameterisations were as follows:

M1	Fixed effects	$\log \rho_{st} = \Psi_{st}$
M2	Spatial heterogeneity	$\log \rho_{st} = \Psi_{st} + \Phi_s$
M3	Spatial clustering	$\log \rho_{st} = \Psi_{st} + \Upsilon_s$
M4	Convolution prior	$\log \rho_{st} = \Psi_{st} + \phi_s + \upsilon_s$
M5	Convolution and month first difference prior	$\log \rho_{st} = \Psi_{st} + \phi_s + \upsilon_s + \omega_{t'(t)}$

Model M1 includes only fixed effects Ψ_{st} in the relative risk $\log \rho_{st}$:

$$
\begin{aligned}
&y_{st}|\mu_{st} \sim \text{NegBin}(\mu_{st}, \kappa) \\
&\log(\mu_{st}) = \log(e_{st}) + \alpha + \delta_{t'(t)} + \sum_j \gamma_j w_{jst} + \sum_j \beta_j x_{jst} \\
&\kappa \sim \text{Ga}(0.5, 0.0005) \\
&\alpha \sim \text{N}(0, 1 \times 10^6).
\end{aligned}
$$

In model M2, a spatially unstructured random effect Φ_s is included, along with the fixed effects Ψ_{st} in the model framework to account for spatial heterogeneity. This is assigned an independent diffuse Gaussian prior $\text{N}(0, \sigma^2_\Phi)$ with a diffuse gamma hyperprior assigned to the precision $\tau_\Phi = 1/\sigma^2_\Phi$ (see Eqn. 5.1):

$$
\begin{aligned}
&y_{st}|\mu_{st} \sim \text{NegBin}(\mu_{st}, \kappa) \\
&\log(\mu_{st}) = \log(e_{st}) + \alpha + \delta_{t'(t)} + \sum_j \gamma_j w_{jst} + \sum_j \beta_j x_{jst} + \Phi_s \\
&\kappa \sim \text{Ga}(0.5, 0.0005) \\
&\alpha \sim \text{N}(0, 1 \times 10^6) \\
&\Phi_s \sim \text{N}(0, \sigma^2_\Phi) \\
&\tau_\Phi \sim \text{Ga}(0.5, 0.0005).
\end{aligned}
$$

In M3 the spatially unstructured random effect Φ_s is replaced by a spatially structured random effect Υ_s, with a diffuse gamma hyperprior assigned to the precision τ_Υ (see

Eqn. 5.3):

$$
\begin{aligned}
&y_{st}|\mu_{st} \sim \text{NegBin}(\mu_{st}, \kappa) \\
&\log(\mu_{st}) = \log(e_{st}) + \alpha + \delta_{t'(t)} + \sum_j \gamma_j w_{jst} + \sum_j \beta_j x_{jst} + \Upsilon_s \\
&\kappa \sim \text{Ga}(0.5, 0.0005) \\
&\alpha \sim \text{U}(-\infty, +\infty) \\
&\Upsilon_s \sim \text{CAR}(\sigma^2_\Upsilon) \\
&\tau_\Upsilon \sim \text{Ga}(0.5, 0.0005).
\end{aligned}
$$

In model M4 the spatial random effect is divided into spatially unstructured ϕ_s and structured components υ_s. The spatially unstructured random effect ϕ_s is assigned an independent diffuse Gaussian exchangeable prior and the structured random effect υ_s is assigned a Gaussian CAR prior (see Eqn. 5.4). As two sources of prior information are assigned to the spatial random effects, their combination is termed the convolution prior:

$$
\begin{aligned}
&y_{st}|\mu_{st} \sim \text{NegBin}(\mu_{st}, \kappa) \\
&\log(\mu_{st}) = \log(e_{st}) + \alpha + \delta_{t'(t)} + \sum_j \gamma_j w_{jst} + \sum_j \beta_j x_{jst} + \phi_s + \upsilon_s \\
&\kappa \sim \text{Ga}(0.5, 0.0005) \\
&\alpha \sim \text{U}(-\infty, +\infty) \\
&\phi_s \sim \text{N}(0, \sigma^2_\phi) \\
&\upsilon_s \sim \text{CAR}(\sigma^2_\upsilon) \\
&\tau_\phi \sim \text{Ga}(0.5, 0.0005) \\
&\tau_\upsilon \sim \text{Ga}(0.5, 0.0005).
\end{aligned} \tag{5.6}
$$

In model M5, a temporal random effect $\omega_{t'(t)}$ is included (see Eqn. 5.5) along with the spatially unstructured ϕ_s and structured υ_s random effects. Month 1 (August) is set to zero ($\omega_1 = 0$) and $\omega_{t'(t)} \sim \text{N}(\omega_{t'(t)-1}, \sigma^2_\omega)$, $(t'(t) = 2, \ldots, 12)$, is assigned to the remaining months. This introduces a random walk into the annual cycle to allow the month effect to depend on the previous month effect. Initially, $\omega_{t'(t)}$ was included in the model framework in addition to the fixed month effect $\delta_{t'(t)}$ contained in Ψ_{st}. However, the inclusion of both the fixed and random month effects resulted in neither of the simulated chains for these parameters converging. Therefore, the fixed effects component of the dengue

relative risk for model M5 omits $\delta_{t'(t)}$ to avoid identifiability problems.

$$
\begin{aligned}
&y_{st}|\mu_{st} \sim \text{NegBin}(\mu_{st}, \kappa) \\
&\log(\mu_{st}) = \log(e_{st}) + \alpha + \sum_j \gamma_j w_{jst} + \sum_j \beta_j x_{jst} + \phi_s + \upsilon_s + \omega_{t'(t)} \\
&\kappa \sim \text{Ga}(0.5, 0.0005) \\
&\alpha \sim \text{U}(-\infty, +\infty) \\
&\phi_s \sim \text{N}(0, \sigma_\phi^2) \\
&\upsilon_s \sim \text{CAR}(\sigma_\upsilon^2) \\
&\omega_{1(t)} = 0,\ \omega_{t'(t)} \sim \text{N}(\omega_{t'(t)-1}, \sigma_\omega^2),\ t'(t) = 2, \ldots, 12 \\
&\tau_\phi \sim \text{Ga}(0.5, 0.0005) \\
&\tau_\upsilon \sim \text{Ga}(0.5, 0.0005) \\
&\tau_\omega \sim \text{Ga}(0.5, 0.0005).
\end{aligned}
$$

For all models, independent diffuse Gaussian priors (mean 0, precision 1×10^{-6}) were taken for the fixed effects β_j $(j = 1, \ldots, 3)$, γ_j $(j = 1, 2)$ and $\delta_{t'(t)}$ with $t'(t) = 2, \ldots, 12$ and a gamma prior was used for the scale parameter κ. The third level of the model is defined so that the variance parameters involved in the second level (random effects) are treated as unknown and given hyperprior distributions. Following Wakefield et al. (2000), weakly informative independent gamma hyperpriors with shape parameter $\zeta = 0.5$ and inverse scale parameter $\eta = 0.0005$ were used for the precisions ($\tau_\phi = 1/\sigma_\phi^2$, $\tau_\upsilon = 1/\sigma_\upsilon^2$, $\tau_\omega = 1/\sigma_\omega^2$) of the priors for the spatially unstructured ϕ_s and structured υ_s random effects $(s = 1, \ldots, 160)$ and temporally autocorrelated random effects $\omega_{t'(t)}$ $(t'(t) = 2, \ldots, 12)$.

5.6 Model implementation

The Bayesian hierarchical models were fitted using WinBUGS software (Lunn et al., 2000) that uses MCMC simulation to produce samples of model parameter values from their joint posterior distribution (see Appendix C for model code). To allow further analyses in R, WinBUGS was called using the **R2WinBUGS** package (Sturtz et al., 2005). Two parallel MCMC chains were generated, each of length 25,000 with a burn-in of 20,000 and thinning of 10 to obtain 1000 samples from the joint posterior distribution

$p(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\mathbf{y})$ (see Appendix B.1). From this, posterior mean estimates of the parameters were calculated (see Appendix B.2). The explanatory climate variables x_{jst} $(j = 1, \ldots, 3)$ and non-climate variables w_{jst} $(j = 1, 2)$ contained within Ψ_{st} (as in Chapter 4) were first standardised to zero mean and unit variance to aid MCMC convergence.

5.6.1 Convergence of Markov chains

The convergence of MCMC chains to a stationary distribution needs be assessed. MCMC samples from the 'log-posterior', i.e, samples from the logarithm of the joint posterior distribution of all model parameters $\boldsymbol{\theta}$, evaluated at each MCMC iteration can be inspected to give an indication of convergence, since the joint posterior distribution is a global summary of all model parameters. Satisfactory convergence of the overall models was confirmed by inspecting the log posterior for the 1000 samples from the joint posterior distribution. The most satisfactory convergence was acheived by model M4 where the log joint posterior distribution is given by $\log\{p(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\gamma}}_j, \hat{\boldsymbol{\phi}}_s, \hat{\boldsymbol{\upsilon}}_s, \hat{\boldsymbol{\tau}}_{\boldsymbol{\phi}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{\upsilon}}|\mathbf{y})\}$, where $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\gamma}}_j, \hat{\boldsymbol{\phi}}_s, \hat{\boldsymbol{\upsilon}}_s, \hat{\boldsymbol{\tau}}_{\boldsymbol{\phi}}, \hat{\boldsymbol{\tau}}_{\boldsymbol{\upsilon}}, \mathbf{y}$ are vectors of length 1000, from the MCMC samples (see Fig 5.1 for model M4). To check convergence of the individual parameter estimates, the potential scale reduction $\hat{R}$ (see Appendix B.3) was calculated for all fixed effects, random effects (unstructured and structured) and hyperparameters (see Fig. 5.2 for model M4). Again, the most satisfactory convergence was found for model M4, where $\hat{R}$ was found to be near to 1 for all estimates (Note: values below 1.1 are thought to be acceptable in most cases, Gelman et al., 2004). Therefore, for this model, the 1000 collected simulations can be treated as samples from the target distribution.

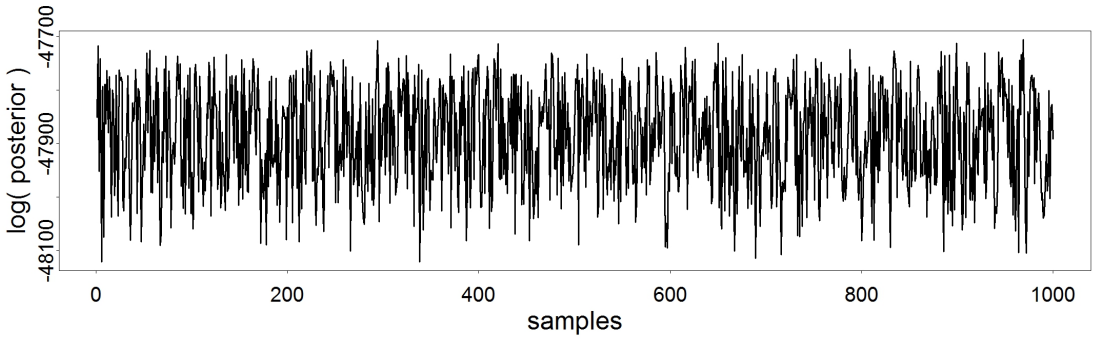


Figure 5.1: Trace plot of log posterior distribution for 1000 samples from model M4.

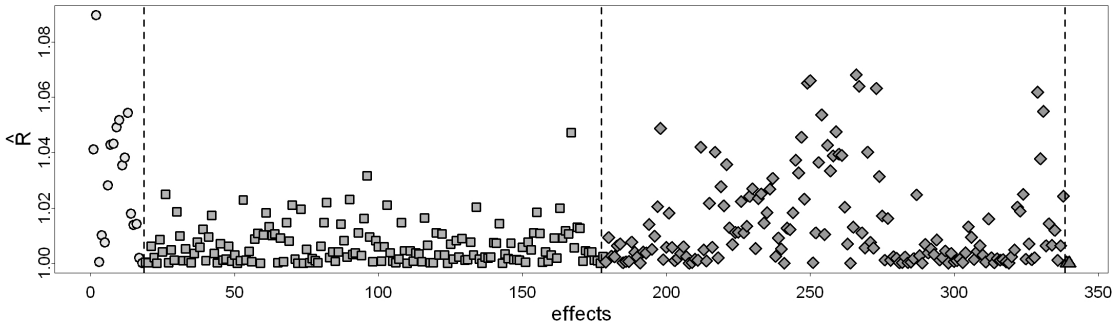


Figure 5.2: Potential scale reduction factor ($\hat{R}$) for 2 chains of 500 samples for fixed effects (circles), unstructured (squares) and structured (diamonds) random effects and hyperparameters (triangles) for model M4. Approximate convergence is diagnosed when $\hat{R}$ is close to 1.

5.6.2 Goodness-of-fit

Attainment of convergence of MCMC algorithms does not necessarily imply good models. A goodness of fit measure widely used in hierarchical Bayesian modelling is the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002). As with the AIC (see Chapter 4, Section 4.4), smaller values of DIC indicate better fitting models (see Appendix B.4). The absolute size of DIC is not relevant, only differences in DIC are important. In general, differences greater than 5 are considered substantial (Spiegelhalter, 2008, see Appendix B.4). Deviance results for each model (M1-M5) are reported in Table 5.1.

Table 5.1: Deviance results for fixed (M1) and mixed (M2-M5) effects models South East Brazil. The posterior mean of the deviance $\bar{D}$, the deviance at the posterior means $D(\hat{\boldsymbol{\theta}})$, the effective number of parameters p_D, the DIC and the overdispersion parameter κ^{-1} (95% credible interval) are given for models M1-M5.

Model	$\bar{D}$	$D(\hat{\boldsymbol{\theta}})$	p_D	DIC	κ^{-1} (95% CI)
M1	100932.3	100914.6	17.7	100950	3.756 (3.674,3.845)
M2	95446.8	95274.8	172	95618.8	2.550 (2.484,2.622)
M3	95447.3	95278.2	169.1	95616.4	2.548 (2.486,2.613)
M4	95446.4	95278	168.4	95614.8	2.549 (2.489,2.612)
M5	95446.5	95278.8	167.6	95614.1	2.547 (2.482,2.611)

For the fixed effects model (M1), the effective number of parameters p_D (see Appendix B.4) is approximately the true number (17) of independent parameters. The inclusion of random effects in the linear predictor caused the estimate of overdispersion κ^{-1} to decrease. This is because random effects allow for extra variation between between different spatial regions. Therefore, the global overdispersion parameter κ^{-1} need only account for residual overdispersion that is not captured by the area-specific random effects. The effective number of parameters is lower for the model which includes the two component convolution prior (M4) compared to models which include an unstructured prior (M2) and a purely spatially structured prior (M3). The inclusion of an unstructured spatial random effect Φ_s in M2 resulted in a substantial reduction in the DIC compared to the fixed effects model (M1) (see Table 5.1).

For comparison, the model was refitted this time replacing the unstructured random effect with a spatially structured random effect that allows for clustering amongst microregions (M3) (see Fig. 5.3.b). Although Figure 5.3a and b are similar, the spatial effect Φ_s in M2 is capturing heterogeneity which is likely attributable to a correlation structure while Υ_s in M3 is designed to allow specifically for local geographical dependence.

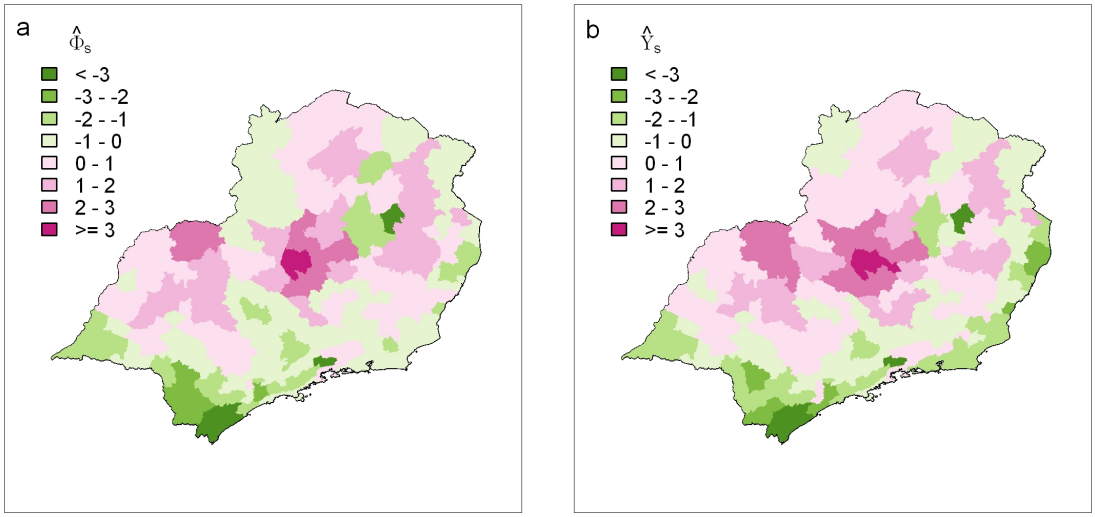


Figure 5.3: Spatial distribution of posterior mean (a) spatially unstructured random effect $\hat{\Phi}_s$, estimated in model M2 and (b) spatially structured random effect $\hat{\Upsilon}_s$, estimated in model M3 for South East Brazil.

In model M4, the spatial random effect was decomposed into the sum of spatially struc-

tured υ_s and unstructured ϕ_s components to account for both clustering between microregions and additional heterogeneity not attributable to similarities between neighbouring areas ($\phi_s + \upsilon_s$). The key feature of the convolution prior is that it allows the assessment of relative contributions of unstructured heterogeneity and spatial clustering to the overall variation of the area effects (MacNab, 2003). Figure 5.4b illustrates that in the South East region, spatial clustering between neighboring areas is the dominant cause of overdispersion. The spatially unstructured random effect ϕ_s in M4 accounts for residual overdispersion in microregions that is not attributable to spatial correlation between the microregions (see Fig 5.4a) and has a minimal contribution to the convolution prior. The combination of the two components of spatial randomness resulted in a small but further reduction in the DIC (Table 5.1).

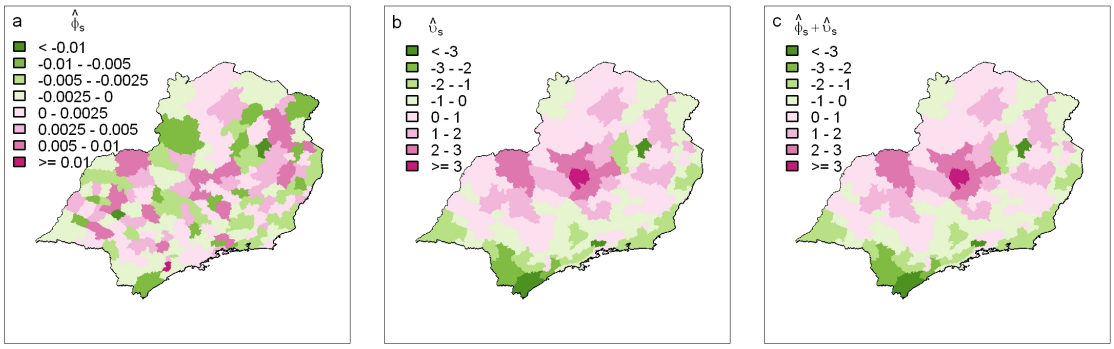


Figure 5.4: Spatial distribution of posterior mean (a) spatial unstructured $\hat{\phi}_s$, (b) structured $\hat{\upsilon}_s$ random effects and (c) their combined effect estimated together in model M4 for South East Brazil.

Following the fit of the model M4, the posterior distribution for the mean $\hat{\boldsymbol{\mu}}_{st}$ can be obtained, where

$$\hat{\boldsymbol{\mu}}_{st} = \exp\left\{\log(e_{st}) + \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\delta}}_{1t'(t)} + \sum_j \hat{\boldsymbol{\gamma}}_j w_{jst} + \sum_j \hat{\boldsymbol{\beta}}_j x_{jst} + \hat{\boldsymbol{\phi}}_s + \hat{\boldsymbol{\upsilon}}_s\right\}.$$

To detect the existence of a temporal dependency structure in the GLMM, deviance residuals from model M4 were estimated based on the posterior mean estimates of the mean $\bar{\mu}$ and the scale parameter $\bar{\kappa}$ using the following formula (from Eqn. 4.3 and Eqn. 4.4)

$$\hat{r}_{(D)st} = sgn(y_{st} - \bar{\mu}_{st})\sqrt{2\left[y_{st}\log(y_{st}/\bar{\mu}_{st}) - (y_{st} + \bar{\kappa})\log\left(\frac{y_i + \bar{\kappa}}{\bar{\mu}_i + \bar{\kappa}}\right)\right]},$$

where $y \log y$ is taken to be zero when $y = 0$. A summary of the distribution of the autocorrelation function of these residuals for the 160 microregions in South East Brazil is shown in Figure 5.5. The deviance residuals show a strong temporal correlation at one month lag and a decreasing correlation up to lags of 5 months.

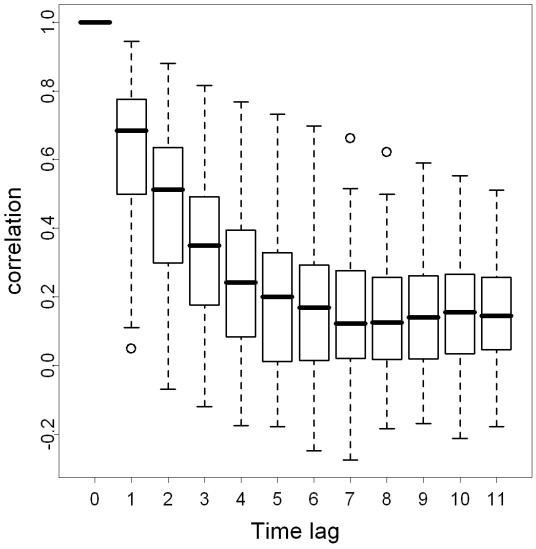


Figure 5.5: Autocorrelation in lagged estimated deviance residuals $\hat{r}_{(D)st}$ across the 160 microregions in South East Brazil (horizontal bar indicates the median, whiskers extend to data point which is no more than 1.5 $\times$ the interquartile range from the box).

In order to allow for this temporal dependence, a random effect $\omega_{t'(t)}$ was introduced. The inclusion of $\omega_{t'(t)}$ in the model framework (M5) resulted in satisfactory convergence. Figure 5.6 shows a comparison of the parameter estimates and 95% credible intervals for the 'fixed' month factor $\delta_{t'(t)}$ in model M4 and the 'autocorrelated' month factor $\omega_{t'(t)}$ in model M5. There is little difference betweeen the two plots. A comparison of DIC for M4 and M5 (see Table 5.1) suggests that the two models are virtually indistinguishable in terms of the overall fit (see Appendix B.4). In light of this and associated convergence diagnostics, the best mixed effects model for dengue relative risk in South Brazil was model M4 (see Eqn 5.6). This model, involving only spatially unstructured and structured random effects, will be used for all subsequent analysis.

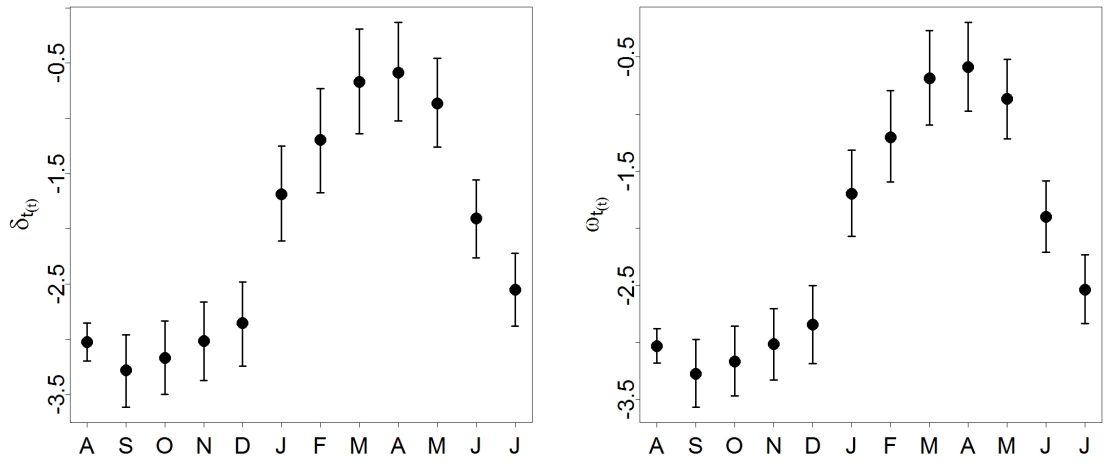


Figure 5.6: Parameter estimates (circle) and 95% credible intervals (bars) for (a) fixed month effect from model M4 and (b) autocorrelated month effect from model M5.

5.6.3 Inference for climate covariates

Posterior densities for the parameters associated with the climate covariates from the selected model, M4, are shown in Figure 5.7. The results for the climate and non-climate covariates and hyperparameters for the spatial random effects are summarised in Table 5.2. For all parameters, the 95% credible interval does not contain zero.

From the GLMM, a 1mm increase in average precipitation over the preceding 3 months in the South East would result in a 39% ($\exp(0.33) = 1.39$, see Table 5.2) increase in dengue relative risk the following month (compared to 9% using the fixed effects GLM, M1). A 1°C increase in temperature over the preceding 3 months would result in a 73% increase in the dengue relative risk (compared to 42% using the fixed effects GLM, M1). Using the fixed effects GLM, an increase in SST anomaly (indicative on El Niño conditions) would result in a 42% decrease in the dengue relative risk estimate for South East Brazil. Using the GLMM, the expected decrease would be 36%. The inclusion of random effects in the model framework allows the climate covariates to ajust to account for variation explained by climate rather than climate and other potentially important unobserved factors. As in Chaper 4, altitude has a negative association, while population density has a positive association with dengue relative risk (see Table 5.2). The reciprocal of the variance

($\tau_\phi = 1/\sigma_\phi, \tau_\nu = 1/\sigma_\nu$) was estimated for the hyperpriors in the hierarchical model (see Table 5.2). Accordingly, the posterior mean variance for the unstructured random effects was found to be very small ($\bar{\sigma}_\phi = 0.00076$, 95% CI=$(0.00021, 0.015)$), while the posterior mean variance of the structured random effects was found to be comparatively larger ($\bar{\sigma}_\nu = 4.202$, 95% CI=$(3.390, 5.348)$). The difference in variability for the unstructured and structured random effects is evident in Figure 5.4a and b respectively (see difference in scale for $\hat{\phi}_s$ and $\hat{\nu}_s$).

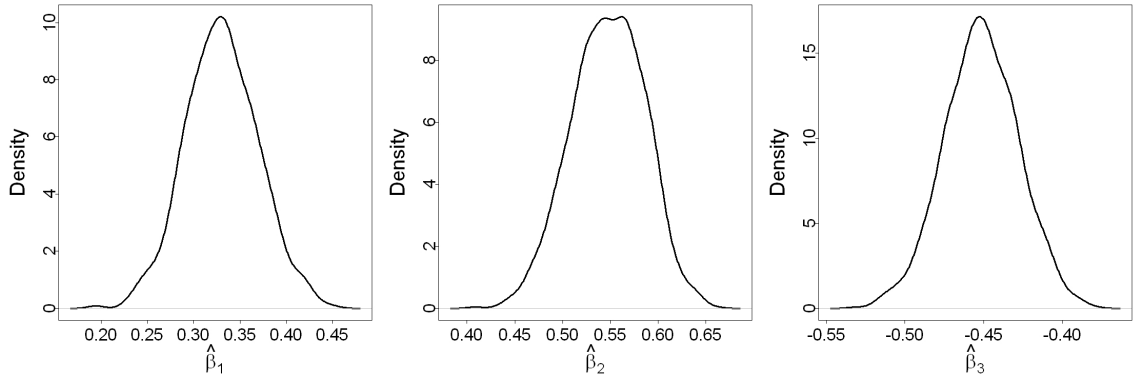


Figure 5.7: Kernel density estimates for the marginal posterior distributions for the parameters $\beta_1, \ldots, \beta_3$ associated with the climate variables (a) average precipitation, (b) average temperature and (c) ONI for South East Brazil, from the negative binomial GLMM.

Table 5.2: Parameter estimates and convergence diagnostic $\hat{R}$ for climate and non-climate covariates and hyperparameters associated with the random effects from negative binomial GLMM, South East Brazil. CI is the credible interval obtained from the 2.5% and 97.5% quantiles of the distribution.

	Prior	Posterior mean (95% CI)	$\hat{R}$
Precipitation	$\beta_1 \sim \text{N}(0, 1 \times 10^6)$	0.33 (0.253,0.412)	1.01
Temperature	$\beta_2 \sim \text{N}(0, 1 \times 10^6)$	0.548 (0.47,0.619)	1.01
Oceanic Niño Index	$\beta_3 \sim \text{N}(0, 1 \times 10^6)$	-0.451 (-0.499, -0.406)	1.00
Altitude	$\gamma_1 \sim \text{N}(0, 1 \times 10^6)$	-1.316 (-1.522, -1.130)	1.09
Population density	$\gamma_2 \sim \text{N}(0, 1 \times 10^6)$	0.214 (0.080, 0.344)	1.00
Unstructured hyperparameter	$\tau_\phi \sim \text{Ga}(0.5, 0.0005)$	1312.099 (64.836, 4697.996)	1.00
Structured hyperparameter	$\tau_\upsilon \sim \text{Ga}(0.5, 0.0005)$	0.238 (0.187, 0.295)	1.00

5.7 Comparison of fixed and mixed effects model

To determine the contribution of random effects to the model fit, results from the selected GLMM were compared to results from the fixed effects GLM. The mean of the posterior distribution of the DIR for all 160 microregions in the South East region were obtained for the 108 month time period (January 2001 - December 2009). Figure 5.8 compares observed and model fit DIR using the GLM (Fig. 5.8a) and GLMM (Fig. 5.8b). More of the variability in dengue cases has been captured by the GLMM, particularly at very low and very high dengue incidence rates. Figure 5.8c presents the time evolution of DIR from January 2001 - December 2009 for South East Brazil. With the exception of the austral summer 2001 when the GLMM overestimates the DIR by a lesser extent than the GLM, the temporal variation of both models is similar.

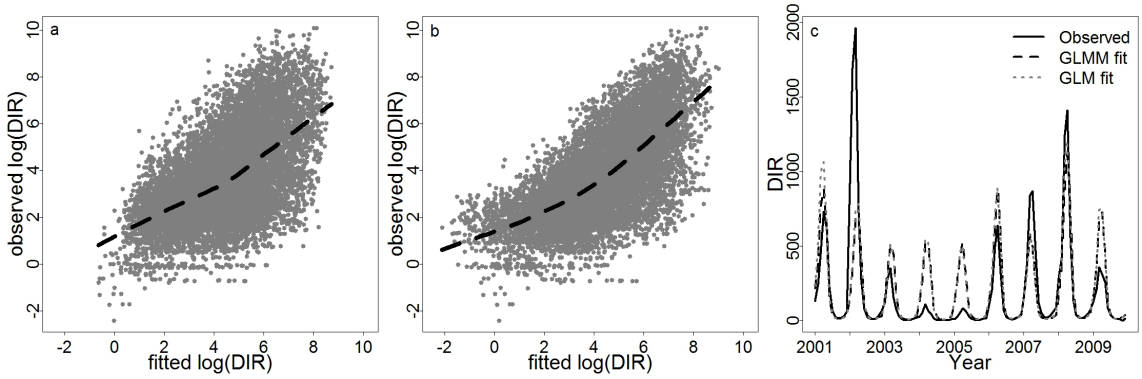


Figure 5.8: Observed and model fit DIR at the linear predictor level for all months (108) and microregions (160) in South East Brazil using (a) GLM and (b) GLMM. Dashed curve - local polynomial regression fit. (c) Total observed (solid line), GLMM (dashed line) and GLM model fit (dotted line) DIR from January 2001 - December 2009.

To assess the ability of the model to capture dengue variability during the peak dengue season, the mean of the posterior distribution of DIR for all 160 microregions in the South East region was extracted for the FMA season. Figure 5.9 compares observed and model fit DIR using the GLM (Fig. 5.9a) and GLMM (Fig. 5.9b) for the FMA season 2001-2009. Again, the GLMM better captures the variability in the DIR for the peak season.

The spatial distribution of observed and model fit DIR for the FMA season from 2001-2009 is presented in Figure 5.10. Although the GLMM does not capture the extent of the

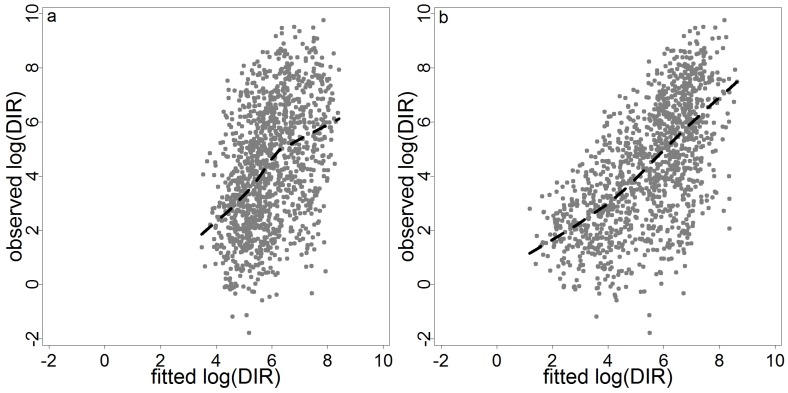
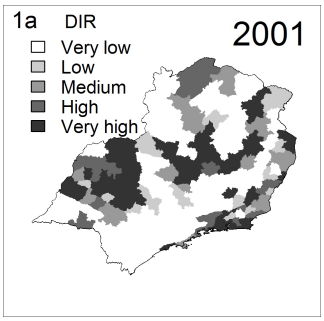
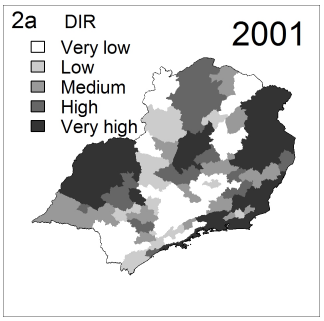
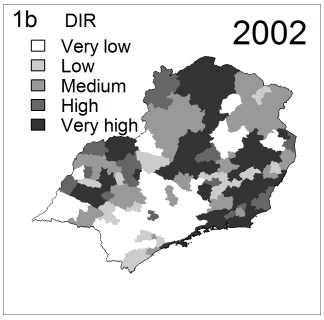


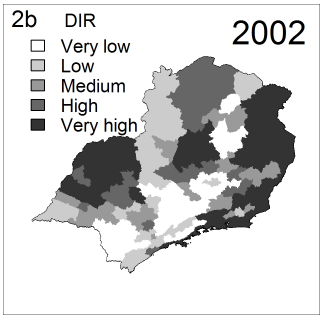
Figure 5.9: Observed and model fit DIR at the linear predictor level for the FMA season 2001-2009 for South East Brazil using (a) GLM and (b) GLMM. Dashed curve - local polynomial regression fit.

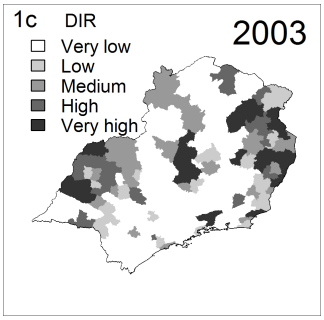
observed inter-annual variability, more instances of high or very high DIR were predicted in FMA 2002 (see Fig. 5.10.1b and 2b) and 2008 (see Fig. 5.10.1h and 2h) compared to 2003-2005 (see Fig. 5.10.1c, 2c, 1d, 2d, 1e, 2e), for example.

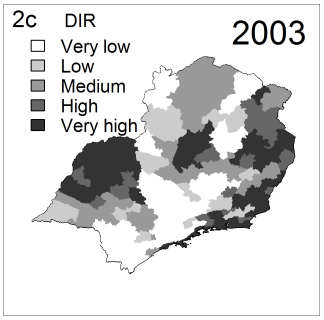


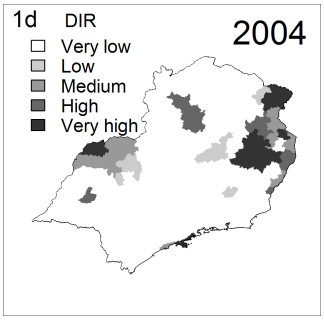


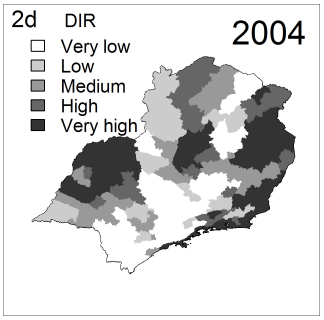


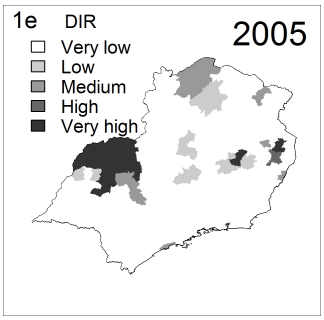


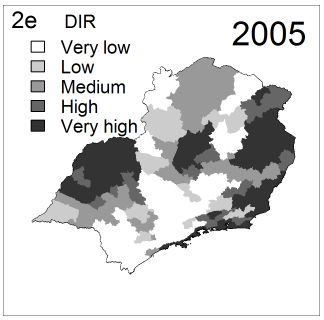












continued overleaf

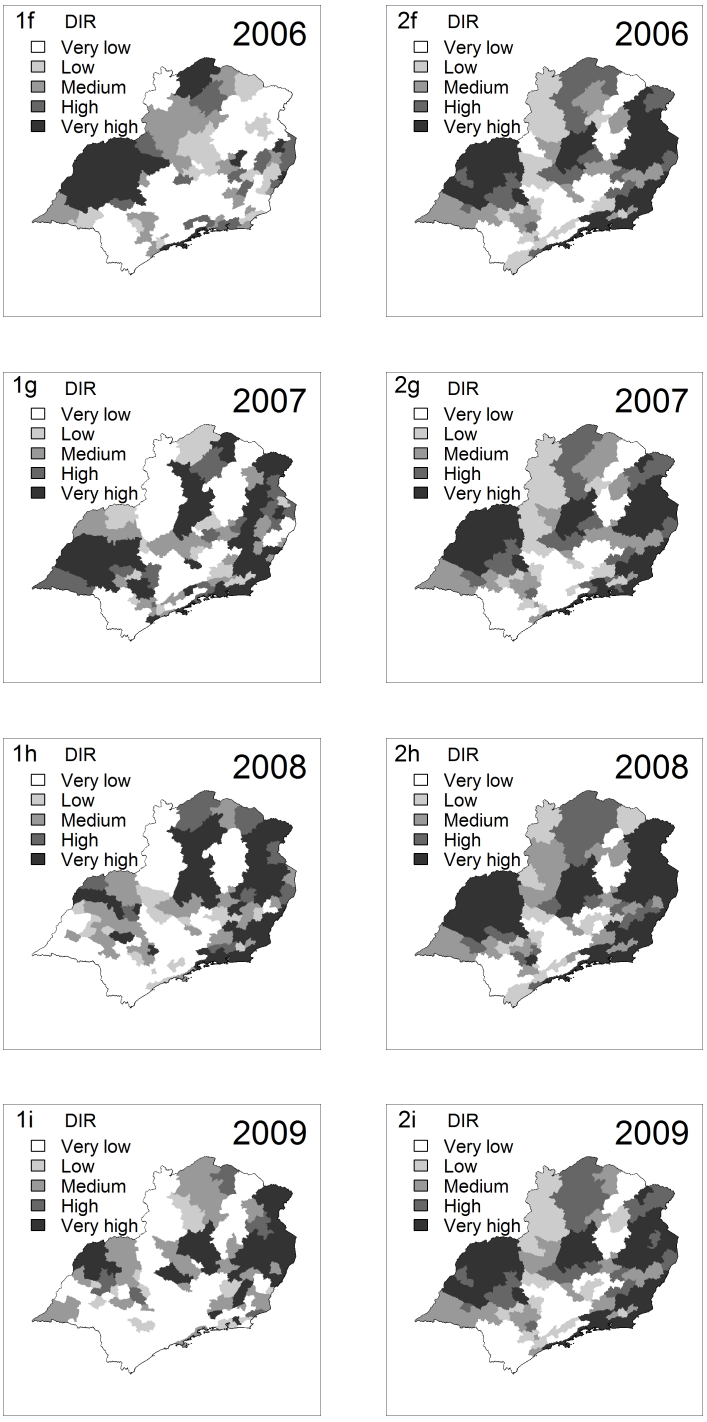


Figure 5.10: Observed (column 1) and model fit (column 2) DIR in South East Brazil for FMA in (a) 2001, (b) 2002, (c) 2003, (d) 2004, (e) 2005, (f) 2006, (g) 2007, (h) 2008 and (i) 2009. Category boundaries defined by 50, 100, 300 and 500 cases per 100,000 inhabitants.

The total DIR at the region level and at an individual microregion level, Rio de Janeiro, are shown in Figure 5.11. Figure 5.11b demonstrates a minimal improvement in the estimated DIR in 2001 and 2004 when using the GLMM. However, both models display an erroneous decrease in the DIR in 2007 and a correct increase and subsequent decrease between 2008-2009. The credible intervals for the GLMM at both the region (Fig. 5.11.1b) and microregion (Fig. 5.11.2b) level are wider, indicative of the variability represented by the spatial random effects in addition to the variability accounted for by the overdispersion parameter κ^{-1}.

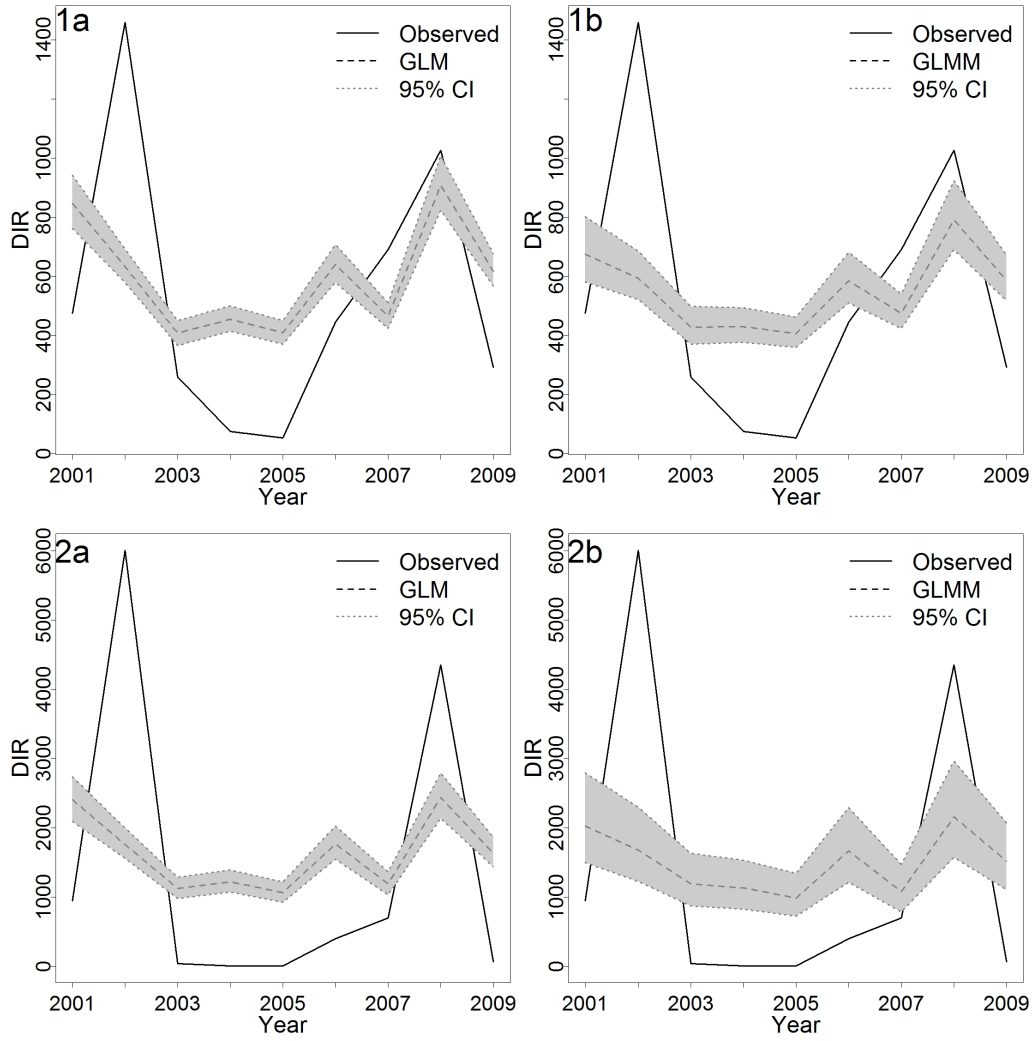


Figure 5.11: Total observed (solid line) and model fit (dashed line) DIR for FMA season 2001-2009 South East Brazil (region level, row 1) and Rio de Janeiro (microregion level, row 2) using (a) GLM and (b) GLMM. Grey shaded area - 95% credible interval.

Although there is little difference in the overall temporal signal between models, there is

a notable improvement in the spatial distribution of DIR due to the inclusion of spatial random effects in the model framework (see Fig. 5.12). The spatial distribution of the observed, GLM fit and GLMM fit DIR in FMA 2008, when a serious epidemic occurred, is shown in Figure 5.12.1a, 1b and 1c respectively. While the GLM provides medium to very high rates across much of the region, the GLMM is able to correctly capture areas with low to very low rates as well as areas with very high rates. Similarly for the FMA season in 2005, a non-epidemic year, the GLMM is able to differentiate between very high DIR in the west of the region and, to some extent, lower rates for the remainder of the region (see Fig. 5.12.2c). These results imply that the inclusion of spatially structured and unstructured random effects allows the model to better capture the variation in DIR across the South East region.

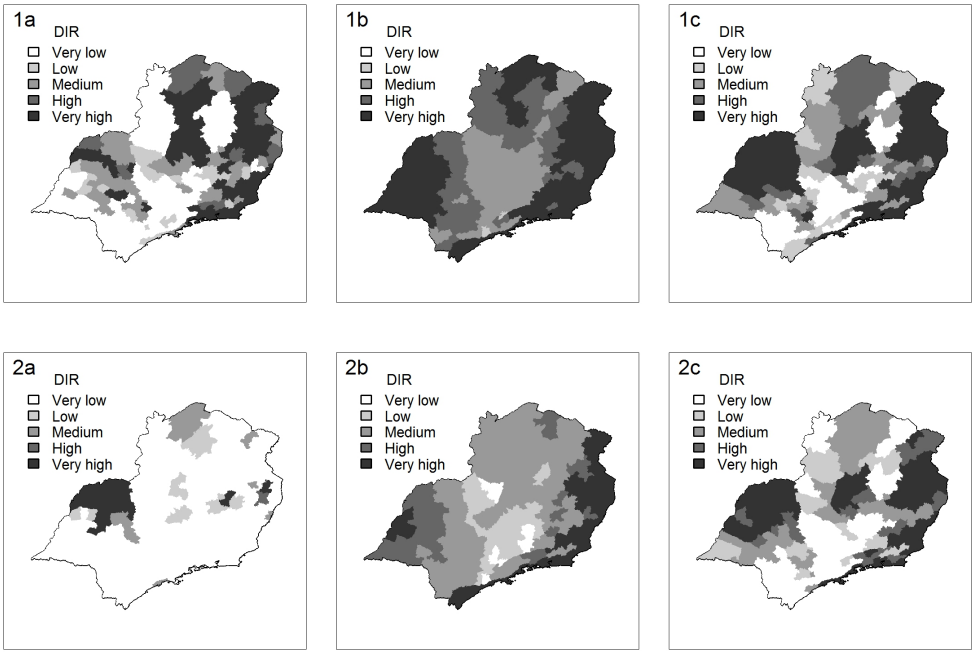


Figure 5.12: (a) Observed, (b) model fit using GLM and (c) model fit using GLMM for South East Brazil, FMA season 2008 (row 1) and 2005 (row 2).

The results thus far suggest that the addition of spatial random effects in the linear predictor improved the overall model fit as confirmed by the DIC and the spatial resolution of modelled dengue incidence rates across South East Brazil. The assignment of a convolution prior model, involving both unstructured independent and spatially structured priors (rather than one or the other) improved convergence of MCMC simulated chains.

The credible intervals for the estimates of the climatic and non-climatic fixed effects did not contain zero. This suggests that these factors play an important role in the transmission of dengue fever in the South East region of Brazil, even when random effects are included in the model formulation. In the following section, the contribution of climate to dengue relative risk will be investigated in more detail.

5.8 Climate contribution to dengue relative risk

Results from Chapter 4 suggested that the ONI influences temporal variations in modelled dengue in the South East region and allowed the model to capture the dengue epidemic in 2008. However, the relationship between ONI and DIR was unexpected. The possibility of leverage and influence was discounted in the previous chapter (Section 4.7) and statistically, ONI appears to be a robust predictor for dengue. In this chapter, ONI remained statistically significant after the inclusion of random effects, which account for unobserved confounding factors. In order to understand the contribution of temperature, precipitation and ONI in predicting dengue relative risk, in both space and time, the contribution of these climatic factors will be explored. This will be done by fitting GLMMs including all three climate covariates, only precipitation and temperature (without ONI) and only ONI (without precipitation and temperature).

5.8.1 Response to ENSO

The selected GLMM was refitted, excluding the ONI index. This resulted in a 'substantial' increase in the DIC (see Appendix B.4) compared to the GLMM including all three climate covariates (see Table 5.3). By including the ONI and excluding precipitation and temperature, the DIC was lower than for the GLMM where ONI was excluded, but greater than for the GLMM where precipitation, temperature and ONI were included.

Figure 5.13 compares the total DIR for the FMA season for the GLMM including all three climate variables (Fig. 5.13a), GLMM excluding ONI (Fig. 5.13b) and GLMM including only ONI (Fig. 5.13c). The GLMM excluding ONI correctly predicts an overall increase in DIR for the region from 2001-2002. Apart from this, the GLMM including all three climate covariates better captures the overall inter-annual variability. Figure 5.14 shows

Table 5.3: Deviance results (posterior mean of the deviance $\bar{D}$, deviance at the posterior means $D(\hat{\boldsymbol{\theta}})$, effective number of parameters p_D, DIC and overdispersion parameter κ^{-1} with 95% credible interval) for GLMM including all three climate covariates, GLMM excluding ONI and GLMM including only ONI.

Model	$\bar{D}$	$D(\hat{\boldsymbol{\theta}})$	p_D	DIC	κ^{-1}
GLMM (precip, temp, ONI)	95446.4	95278	168.4	95614.8	2.549 (2.489,2.612)
GLMM (precip, temp)	95795.7	95629.1	166.5	95962.2	2.615 (2.550,2.681)
GLMM (ONI)	95706.7	95541.1	165.6	95872.3	2.604 (2.537,2.674)

the decomposition of the dengue relative risk across the South East into the climate components from the GLMM including all three climate variables (Eqn. 5.7) and the GLMM excluding ONI (Eqn. 5.8).

$$\text{SMR}_{st} = \exp(\beta_1 x_{1st} + \beta_2 x_{2st} + \beta_3 x_{3t}) \tag{5.7}$$

$$\text{SMR}_{st} = \exp(\beta_1 x_{1st} + \beta_2 x_{2st}) \tag{5.8}$$

It is clear that including ONI allows more of the spatial and temporal variability to be captured. The ONI varies only in time and contains no spatial information. However, including the ONI in the model, along with precipitation, temperature and spatially unstructured and structured random effects, allows the area specific random effects to adjust to the uneven impact that ENSO may have on dengue relative risk in different areas.

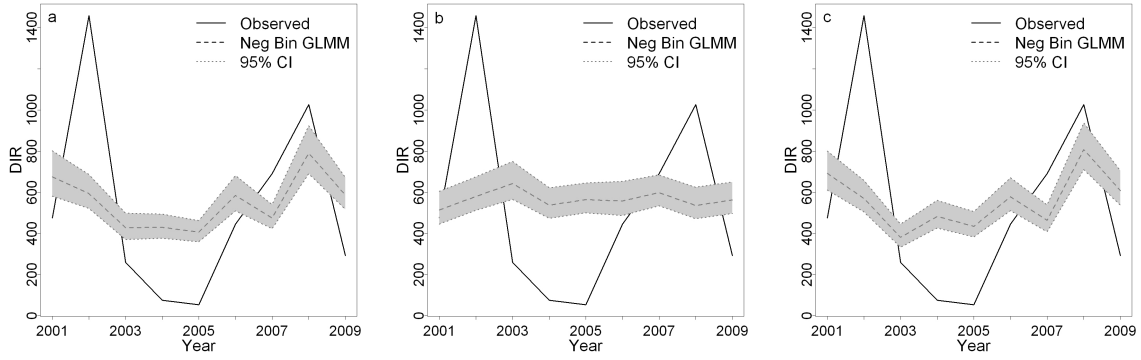
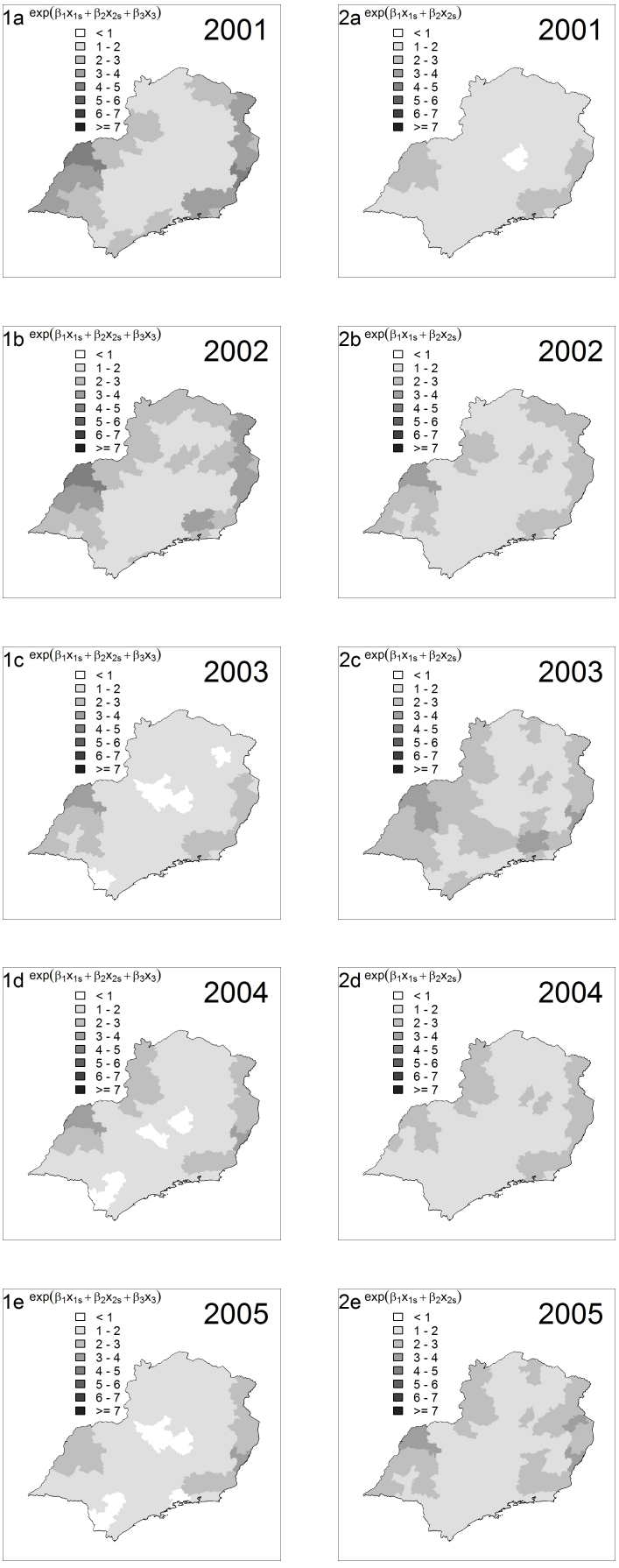


Figure 5.13: Total observed (solid line) and model fit (dashed line) DIR for FMA season 2001-2009 South East Brazil using GLMM with (a) all 3 climate variables, (b) excluding ONI and (c) Including only ONI. Grey shaded area - 95% credible interval.



continued overleaf

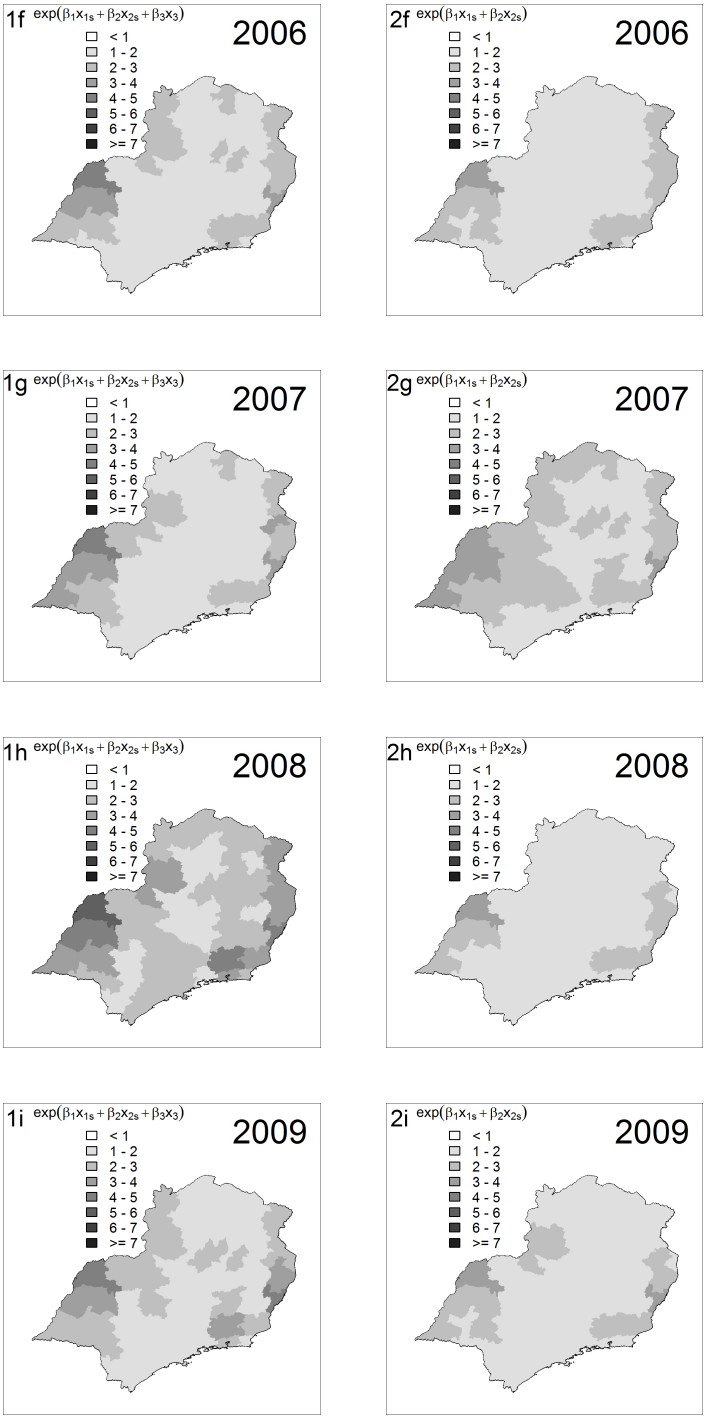


Figure 5.14: Multiplicative decomposition of the dengue relative risk map in South East Brazil into the climate component explained by precipitation, temperature and ONI (Eqn. 5.7) from GLMM including all three variables (column 1) and precipitation and temperature (Eqn. 5.8) for GLMM excluding ONI (column 2) for FMA season in (a) 2001, (b) 2002, (c) 2003, (d) 2004, (e) 2005, (f) 2006, (g) 2007, (h) 2008 and (i) 2009. The decomposition is in terms of posterior means.

5.8.2 Response to local climate for different ENSO scenarios

Collinearity between the three climate explanatory variables may cause the true relationship of each to be misrepresented. Each posterior estimate may borrow strength from other correlated variables. However, a model with correlated variables does not reduce the predictive power of the overall model. In order to understand how the climate explanatory variables predict dengue relative risk, the climate contribution to dengue relative risk (Eqn. 5.7), under a range of scenarios can be determined. For example, for varying precipitation rates and temperatures under El Niño, Neutral and La Niña conditions (see Chapter 3, Section 3.4.4). Figure 5.15 illustrates the climate contribution to dengue relative risk given three different ENSO conditions: La Niña (e.g. ONI= -1), Neutral (e.g. ONI= 0) and El Niño (e.g. ONI= $+1$). The range (maximum - minimum) of the standardised values of precipitation and temperature were determined for the DJF season 2000-2009. In general, as both precipitation and temperature increase, so does dengue relative risk. However, according to the model, dengue relative risk increases at a greater rate given a La Niña event (Fig. 5.15a). Previous results (see Chapter 3) suggest that typical 3 month values of precipitation and temperature are more likely to occur towards the bottom left corner of the dengue relative risk space given La Niña conditions, but towards the top right corner given El Niño conditions. It is possible that there is some optimum average precipitation and temperature combination which is most conducive to increased levels of dengue relative risk. This optimum level may be achieved during La Niña conditions which results in precipitation and temperature levels most suitable for the mosquito and dengue transmission. Although precipitation and temperature have an overall positive effect on dengue relative risk, ENSO may act as an indicator that these two variables have reached an optimum state for increased dengue incidence. However, this is highly speculative and without further research as to the biological implications of precipitation and temperature thresholds for dengue and or the availability of a longer time series, it is difficult to reach conclusions as to the role of ENSO in predicting dengue epidemics in South East Brazil.

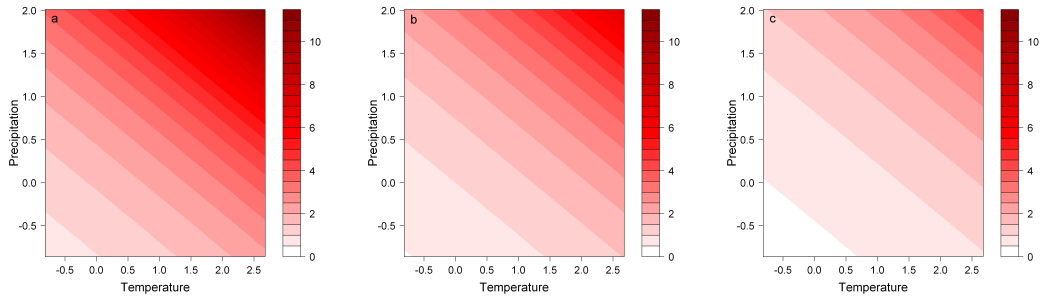


Figure 5.15: Climate contribution, SMR $= \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$, to dengue relative risk for DJF precipitation $(\min(x_1) < x_1 < \max(x_1))$ and DJF temperature $(\min(x_2) < x_2 < \max(x_2))$ combinations under (a) La Niña ($x_3 = -1$), (b) Neutral ($x_3 = 0$) and (c) El Niño ($x_3 = 1$) conditions.

5.9 Conclusion

In this chapter, a GLMM was tested and implemented via a Bayesian framework using MCMC, to better capture spatio-temporal variations in dengue relative risk. The addition of spatial random effects in the model framework captured spatial heterogeneity between microregions due to unknown/unobserved confounding factors and spatial clustering between microregions. This improved the ability of the model to spatially capture variability in dengue incidence rates across South East Brazil. Precipitation, temperature and ONI are statistically significant predictors of dengue. However, the role of ENSO may be obscured by local climate heterogeneity, insufficient data, randomly coincident outbreaks, and other, potentially stronger, intrinsic factors regulating transmission dynamics (Johansson et al., 2009a). Using a Bayesian approach, posterior predictive distributions for disease risk can be derived at each spatial location for a given month or season. This allows probabilistic forecasts to be issued, which is useful for developing disease early warning systems as forecast uncertainty can be quantified. Although climate information alone does not account for a large proportion of the overall variation in dengue cases in Brazil, spatio-temporal climate information with the addition of spatial random effects do account for some of this variability, particularly for the 2008 peak dengue season in South East Brazil, when a serious epidemic occurred. Therefore, the possibly of a climate driven dengue early warning for Brazil is worth investigating. The subsequent chapter will investigate the predictive validity of the model to provide probabilistic forecasts of future and geographically specific dengue epidemics.

Chapter 6

Towards a dengue early warning system for South East Brazil

6.1 Introduction

The focus of the previous chapters was on model selection and development using all available data, to gain a scientific understanding of the extent and significance of climate-dengue relationships. However, as outlined in Chapter 1, it is important to assess how well the developed model can predict future and also geographically specific dengue epidemics. Therefore, the focus of this chapter shifts to prediction. The aim is to determine the best dengue prediction model, given the available information, and to test the ability of these models to predict future dengue epidemics across the South East of Brazil. The GLMM developed in Chapter 5 is compared to a simpler approach representative of the current practice in dengue surveillance in Brazil, based on 3-month persistence of dengue relative risk. To assess how well the models perform in prediction mode, both models are fitted to data from April 2001 - December 2007 and tested on out-of-sample data from January 2008 - December 2009. The potential benefit of combining the newly developed GLMM with a formalised model of current practice is then considered. Several prediction tools are used including posterior predictive distributions and a novel technique for visualising probabilistic forecasts. The warnings from the forecasting systems are evaluated using Receiver Operating Characteristic (ROC) analysis. This investigation helps to quantify the predictive benefit of the model and to ensure the efficacy of the modelling framework

to public health decision makers.

6.2 Mathematical model based on current practice

As discussed in Chapter 1, the current monitoring system in Brazil relies on observing an increase in early cases around 3 months prior to the onset of the peak dengue season. To test if the developed spatio-temporal hierarchical model performs better than current practice, the GLMM will be compared to a simple mathematical model based on current practice:

$$y_{st}|\mu_{st} \sim \text{NegBin}(\mu_{st}, \kappa)$$
$$\log \mu_{st} = \log e_{st} + \alpha + \gamma_0 \log(\tfrac{y_{st-3}}{e_{st-3}}),$$

with the expected number of cases e_{st} (see Chapter 3, Section 3.2.3, Eqn. 3.2) as the model offset. The variable (y_{st-3}/e_{st-3}) is the observed divided by the expected cases (i.e. the SMR) lagged by 3 months. This lag was selected as a compromise between the longest lag plausible to provide predictive skill and the shortest lag possible to allow enough time to provide an early warning of a dengue epidemic. For example, a dengue prediction for March 2011 would be based on the SMR for December 2011. The model was fitted to the dengue data (April 2001- December 2009) using WinBUGS (see Chapter 5, Section 5.2). The inclusion of an autoregressive term causes the first 3 observations in each microregion to be lost. To perform a fair comparison of this current practice autoregressive model (ARM) to the model developed in Chapter 5, the GLMM (Eqn 5.6) was refitted excluding the first 3 months of the dataset (January - March 2001). Table 6.1 presents deviance results for the two models.

Table 6.1: Deviance results for GLMM and ARM (representative of current practice) for South East Brazil.

Model	$\bar{D}$	$D(\hat{\boldsymbol{\theta}})$	p_D	DIC	κ^{-1} (95% CI)
GLMM	91856.2	91689.7	166.5	92022.7	2.531 (2.468, 2.597)
ARM	102854.1	102851.2	2.907	102857.0	5.565 (5.444, 5.698)

The DIC is considerably greater for the ARM, suggesting that the GLMM is a better fit. The overdispersion parameter κ^{-1} is also greater for the ARM suggesting that using

past dengue relative risk as a model covariate captures a lot less variability in the dengue data than using climate and non-climate covariates and random effects.

Figure 6.1 shows the dengue incidence rate (DIR) for the South East region using the GLMM and the ARM from April 2001 - December 2009. Although the ARM was able to detect low DIR in 2004 and 2005, any warning for the epidemic years 2002 and 2008 would have arrived after the epidemic peaks, which would be too late to implement effective interventions (see Fig. 6.1b).

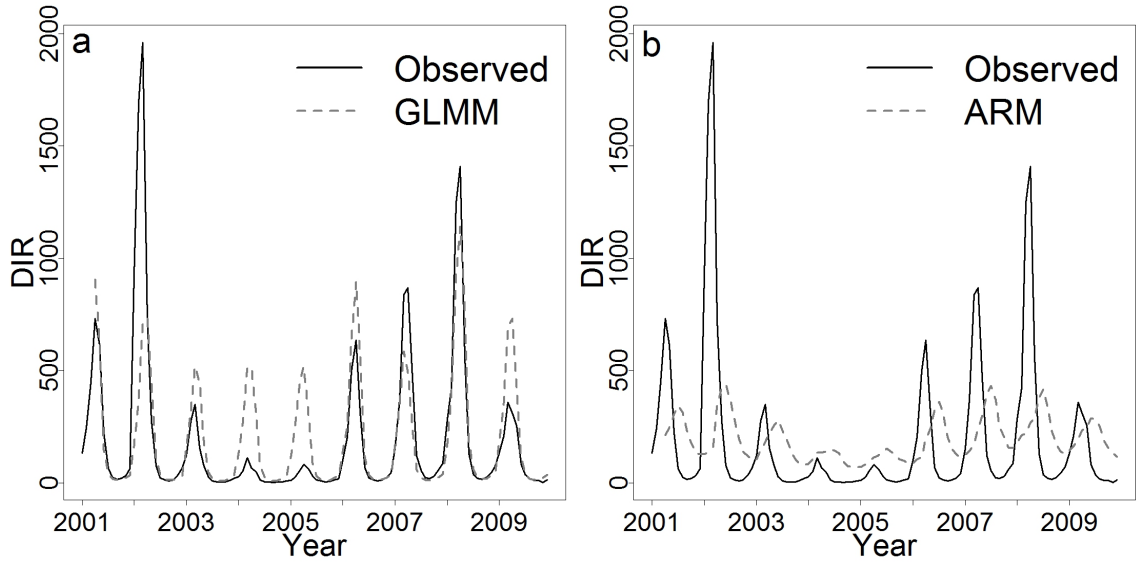


Figure 6.1: Total observed (solid line) and model fit (dashed line) DIR from April 2001 - December 2009 for (a) GLMM and (b) ARM.

Figure 6.2 shows the spatial distribution of observed and model fit DIR using the GLMM and the ARM for the FMA season 2002 - 2009 (note: due to the use of a lagged term, the first season 2001 is lost). Although the GLMM has a tendency to over predict DIR in certain areas, the model is better able to capture instances of very high DIR across the South East region. The ARM better captures low levels of DIR across the region in 2004 and 2005 (Fig. 6.2.3c and 3d). However, in general the ARM predicts low to medium DIR for most of the region even when high DIR is observed. Despite some false alarms (i.e. high DIR predicted when low DIR observed), there are more instances where the GLMM successfully detected high DIR compared to the ARM (e.g. west of the region 2004, Fig. 6.2.2d, east coast 2008, Fig. 6.2.2g).

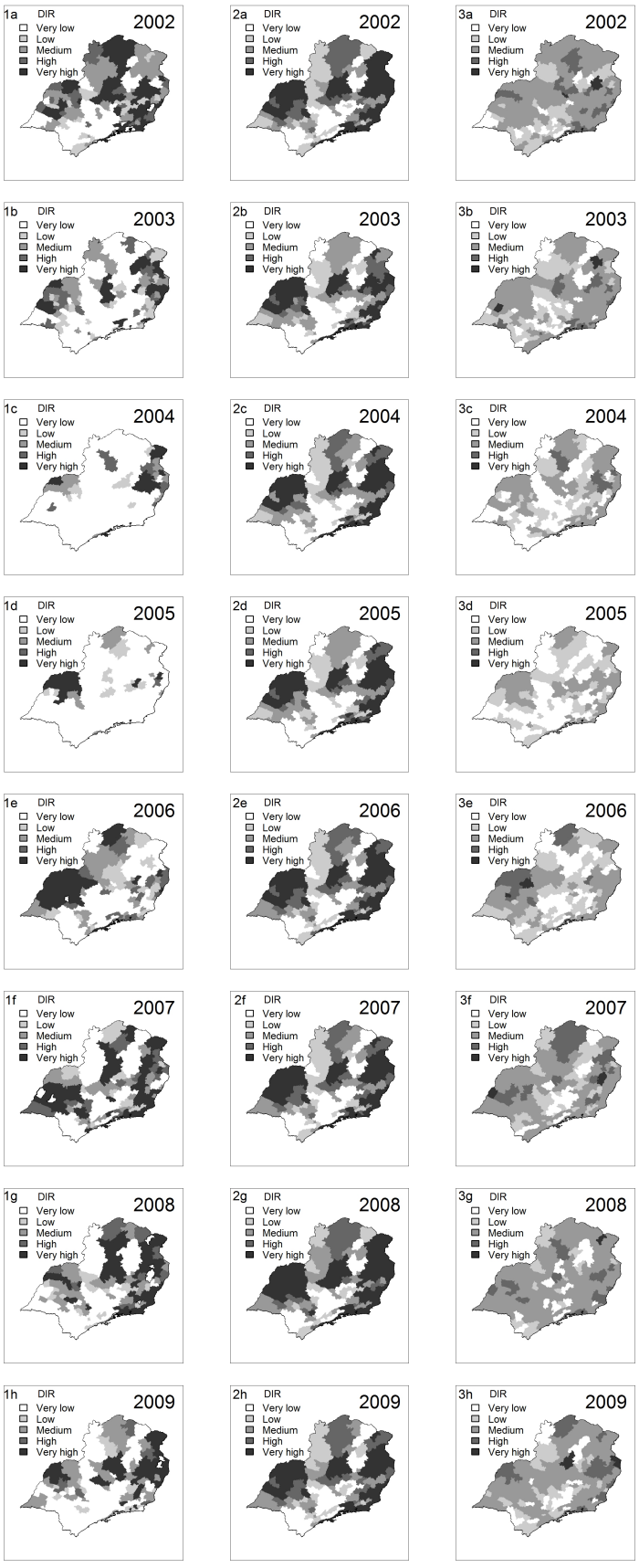


Figure 6.2: Observed DIR (column 1), model fit DIR using GLMM (column 2) and autoregressive model (column 3) for FMA season in (a) 2002, (b) 2003, (c) 2004, (d) 2005, (e) 2006, (f) 2007, (g) 2008 and (h) 2009. Category boundaries defined by 50, 100, 300 and 500 cases per 100,000 inhabitants.

6.3 Posterior predictive distributions

The previous section indicates that the GLMM can be argued to fit the data better. However, it is also useful to look at the predictive ability of both models to ascertain if the GLMM is capable of issuing better probabilistic warnings than the model based on current practice. When assessing complex Bayesian models, it can be useful to use posterior predictive distributions as reference distributions for comparison to observed data (Gelman et al., 1996). The posterior predictive distribution of the response is obtained by simulating new pseudo-observations using samples from the posterior distribution of the parameters in the model (see Appendix B.5). The distribution of estimated values can then be compared to observed values.

For example, consider the GLMM specified in the previous chapter, where given the random effects, the data distribution is negative binomial:

$$\begin{aligned} \tilde{\mathbf{y}}_{st}|\hat{\boldsymbol{\mu}}_{st} &\sim \text{NegBin}(\hat{\boldsymbol{\mu}}_{st}, \hat{\boldsymbol{\kappa}}) \\ \hat{\boldsymbol{\mu}}_{st} &= \exp\left\{\log(e_{st}) + \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\delta}}_{t'(t)} + \sum_j \hat{\boldsymbol{\gamma}}_j w_{jst} + \sum_j \hat{\boldsymbol{\beta}}_j x_{jst} + \hat{\boldsymbol{\phi}}_s + \hat{\boldsymbol{\upsilon}}_s\right\}, \end{aligned}$$

where $\tilde{\mathbf{y}}_{st}$ is a vector of posterior predictions of unobserved future dengue counts, $\hat{\boldsymbol{\mu}}_{st}$ is a vector of posterior estimates of the mean dengue counts and $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\gamma}}_j, \hat{\boldsymbol{\delta}}_{t'(t)}, \hat{\boldsymbol{\phi}}_s, \hat{\boldsymbol{\upsilon}}_s, \hat{\boldsymbol{\kappa}}$ are vectors of length 1000, obtained from the MCMC samples (see Section 5.2), from a model fitted to the first 7 years of data (January 2001 - December 2007). Using these parameter estimates and the climate (x_{jst}) and non-climate (w_{jst}) explanatory variables from January 2008 - December 2009, the posterior distribution of the mean dengue count $\hat{\boldsymbol{\mu}}_{st}$ can be calculated. Subsequently, the posterior predictive distribution of dengue counts, $\tilde{\mathbf{y}}_{st}$, is estimated by drawing random values from a negative binomial distribution with mean corresponding to the elements of $\hat{\boldsymbol{\mu}}_{st}$ and scale parameter corresponding to the elements of $\hat{\boldsymbol{\kappa}}$, estimated from the model. Figure 6.3 shows a comparison of the posterior distribution $\hat{\boldsymbol{\mu}}_{st}$ and the posterior predictive distribution $\tilde{\mathbf{y}}_{st}$ of dengue counts for the microregion of Rio de Janeiro for March 2008. As the posterior predictive distribution is simulated from a negative binomial distribution with a scale parameter that was estimated to be close to zero (mean of the posterior distribution $\bar{\kappa} = 0.36$), the resulting posterior predictive distribution is highly positively skewed.

The predictive ability of the GLMM and ARM was tested by obtaining the posterior

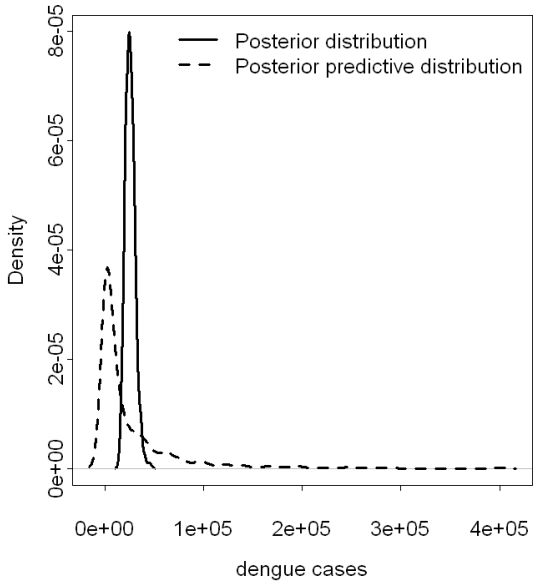


Figure 6.3: Posterior (solid line) and posterior predictive (dashed line) probability densities of dengue counts for microregion of Rio de Janeiro for March 2008.

predictive distributions of $\tilde{\mathbf{y}}_{st}$ for each microregion from January 2008 - December 2009. Random samples from a negative binomial distribution with mean equal to the MCMC samples were obtained from both models fitted to April 2001 - December 2007. The predictions for January 2008 - December 2009 were compared with observations for each of the 160 microregions in South East Brazil. In general, dengue warnings are most useful at the microregion level, to allow local governments to make decisions on resource allocation. Because of this, it is useful to consider a selection of microregions for further inspection. Five microregions were selected as follows (see Fig. 6.4). Rio de Janeiro (population of 11,554,872) and Belo Horizonte (population of 4,932,777) were chosen as they contain the capital cities of the states of Rio de Janeiro and Minas Gerais, respectively. As São Paulo experienced comparatively low DIR during the out-of-sample period, another large microregion in the state of São Paulo was selected: São Jose dos Campos (population of 1,381,846). Três Marias (population of 97,652), situated in northwestern Minas Gerais was selected as one of several microregions in this area where the GLMM performed relatively well. Baía de Ilha Grande (population of 199,373) is a popular holiday resort in the state of Rio de Janeiro. Dengue early warnings for this location could help national and international tourists prepare to protect themselves from the dengue mosquito.

In Figure 6.5, observed DIR, the mean of the posterior predictive distribution and 95%

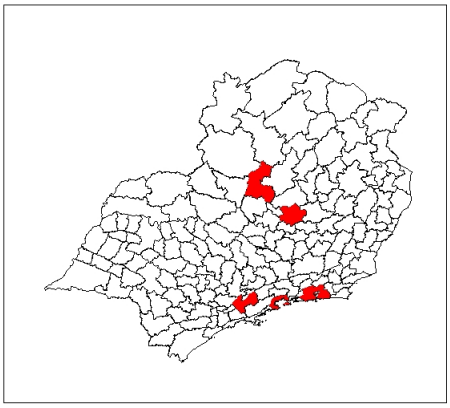


Figure 6.4: Location of five selected microregions (clockwise from most northern microregion) Três Marias, Belo Horizonte, Rio de Janeiro, Baía de Ilha Grande and São Jose dos Campos.

credible intervals, calculated using the 2.5% and 97.5% quantiles of the posterior predictive distribution, are presented for these five microregions. In general, the GLMM better captured the seasonality of DIR. In the microregion Três Marias, the GLMM successfully predicted an increase in DIR in the peak dengue season in 2008 and a lower DIR for the following season in 2009 (see Fig. 6.5.1a). The GLMM was also able to predict that the dengue season for Belo Horizonte was equally high in 2009 as in 2008 (see Fig. 6.5.1b). Using the ARM, the peak DIR in 2008 was not captured by the 95% credible interval for the microregion Baía de Ilha Grande, whereas the 95% credible interval for the GLMM narrowly encompassed the peak in DIR (see Fig. 6.5.1c and 6.5.2c). For microregions Rio de Janeiro and São Jose dos Campos the GLMM slightly over-predicted the 2009 season but better captured the behaviour in dengue than the ARM (see Fig. 6.5.1d, 6.5.2d, 6.5.1e and 6.5.2e). As the scale parameter $\hat{\boldsymbol{\kappa}}$ was lower for the ARM (fitted to the 2001-2007 data) than for the GLMM ($\bar{\kappa} = 0.17$ for ARM, $\bar{\kappa} = 0.36$ for GLMM), credible intervals for the ARM model are considerably larger.

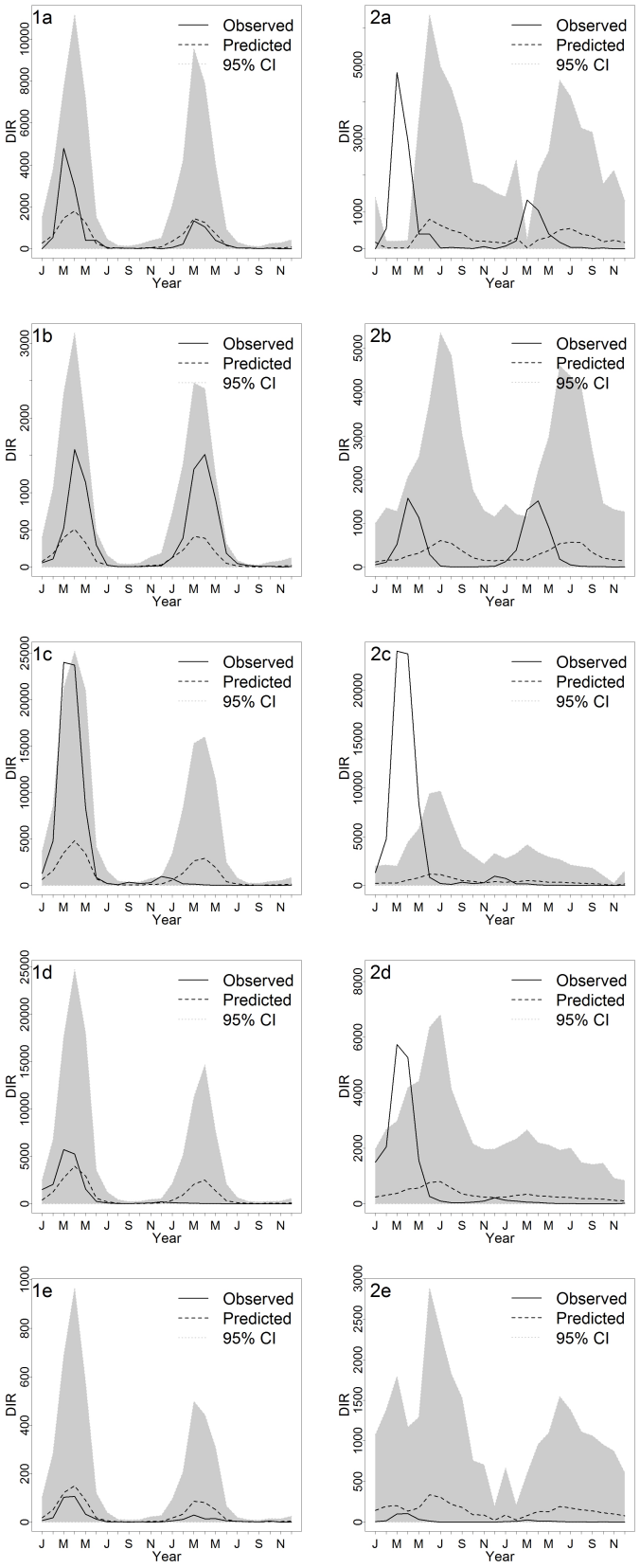


Figure 6.5: Time series of observed (solid line), posterior predictive mean (dashed line) and 95% credible intervals for posterior predictive distribution from January 2008 - December 2009 using GLMM (column 1) and ARM (column 2) for selected microregions: (a) Três Marias, (b) Belo Horizonte, (c) Baía de Ilha Grande, (d) Rio de Janeiro and (e) São Jose dos Campos. Note different scales.

6.4 Visualising ternary probabilistic forecasts

Although posterior predictive distributions for individual microregions could be of use to local decision makers, in practice a summary of the information contained in posterior predictive distributions for each microregion in the form of a map may be more useful for targeting resource allocation to areas most at risk. As posterior predictive distributions for dengue incidence rates can be derived for each microregion and month, the probability of dengue risk falling into pre-defined categories can be calculated. Summarising forecast information into categories may be a better way of communicating probabilistic warnings of dengue risk to public health decision makers. A probabilistic forecast consists of a set of probabilities assigned to possible outcomes. Here, attention will be restricted to ternary forecasts that assign probabilities to a set of three mutually exclusive and complete outcomes (e.g. low, intermediate and high risk). Category boundaries can be defined in an appropriate way for the application and do not have to be evenly spaced. Seasonal climate forecasts are commonly issued in terms of tercile categories (e.g. Barnston et al., 2003; Palmer et al., 2004; Saha et al., 2006). In this case, terciles of the historical observational dataset at each spatial location define the boundaries of the categories. However, public health decision makers may prefer to use pre-defined categories which are more meaningful to the disease in question. The labels B, N, and A will be used to denote 'below normal', 'near normal', and 'above normal' values of the forecast variable. When using terciles, one third of the historical observations lie in each of the categories B, N, and A. Given these categories, the forecasting system can produce probabilistic forecasts, p_B, p_N, p_A, that a variable, e.g. y_{st}, will be in each category at the forecast time. The probability forecast can be regarded as $\mathbf{p} = (p_B, p_N, p_A)$ with the constraints $p_B + p_N + p_A = 1$ and $0 \leq p_i \leq 1$, $\forall i$. The particular forecast $\mathbf{q} = (q_B, q_N, q_A)$ corresponds to the case where the forecaster's state of knowledge is 'no better' than the historical observed distribution of $\mathbf{y}$. For example, if the forecaster had no knowledge other than the observational record, the same forecast $\mathbf{q}$ could be issued each year. Here, $\mathbf{q}$ will be referred to as the reference forecast. In climate science this distribution is referred to as *climatology*. The reference forecast can be viewed as a benchmark distribution with which all other forecasts can be compared.

If a 'forecasting system' is capable of producing probabilistic forecasts over a geographical area, these forecasts can be displayed graphically in the form of a map. To communicate

information contained in a probabilistic forecast, a new method for visualising ternary probabilistic forecasts will be adopted. This method is described in Appendix D and a more detailed account can be found in a manuscript in preparation entitled 'On the interpretation, verification and calibration of ternary probabilistic forecasts', by Jupp, T.E., Lowe, R., Stephenson D.B. and Coelho, C. A. S (Jupp et al., 2010). The idea is to consider a ternary forecast as a point in a triangle of barycentric coordinates. This allows a unique colour to be assigned to each forecast from a continuum of colours defined on the triangle. Colour saturation increases with information gain relative to the reference forecast (the difference between the forecast $\mathbf{p}$ and the reference forecast $\mathbf{q}$, see Appendix D, Section D.5). This provides additional information to decision makers compared to conventional methods used in seasonal climate forecasting, where one colour is used to represent one forecast category on a forecast map (e.g. red='dry', see Appendix D, Section D.3, Fig D.2).

Using this new method, maps can be produced in which the forecast at each geographical location is expressed as a colour determined by a combination of three probabilities. In terms of dengue prediction, it is possible to use tercile boundaries of the historical observations to define the three categories. However, for epidemic decision making purposes, alternative pre-defined risk boundaries might be more meaningful than terciles. For example, as described in Chapter 3, Section 3.2.2, the Brazilian Ministry of Health are interested in areas where the DIR $\leq$ 100 cases per 100,000 inhabitants; indicating low risk, 100 $<$ DIR $\leq$ 300 cases per 100,000 inhabitants; indicating medium risk and DIR $>$ 300 cases per 100,000 inhabitants; indicating high risk.

According to the observed distribution for the FMA season 2001-2007, 65% of the values fell below DIR = 100, 12% fell between DIR = 100 and DIR = 300, and 23% fell above DIR = 300. As the categories apply to a dengue rate, rather than absolute counts, the category boundaries are the same for each spatial location. Therefore, the reference forecast $\mathbf{q}$ becomes $\mathbf{q} = (0.65, 0.12, 0.23)$, rather than $\mathbf{q} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ for evenly spaced tercile boundaries (see Fig. 6.6). When representing probabilistic forecasts using colour, determined from a point in a triangle of barycentric coordinates (see Appendix D, Section D.5), the reference forecast ($\times$) shifts from the centre of the triangle (for category boundaries defined by terciles, see Fig. 6.6a) to a point which satisfies these 3 probabilities (Fig. 6.6b). Using these category boundaries, blue is assigned to category B (low risk), yellow to category N (intermediate risk) and red to category A (high risk). In Figure 6.6b,

the white area of the triangle, which depicts a forecast when $\mathbf{p} \approx \mathbf{q}$, shifts towards the bottom left causing the colour scheme to differ from that of tercile category boundaries.

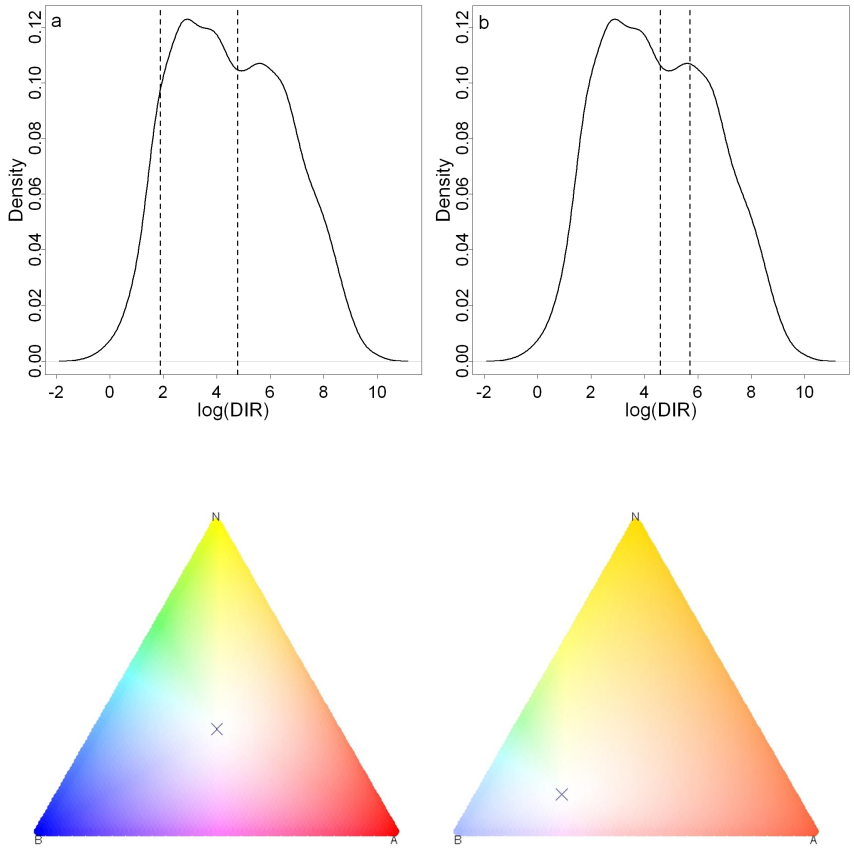


Figure 6.6: Kernel density of FMA DIR in South East Brazil 2001-2007 with (a) tercile category boundaries (dashed lines) and (b) pre-defined category boundaries (dashed lines) of 100 and 300 cases per 100,000 inhabitants. Note logarithmic scale. Labels B, N, and A denote 'below normal', 'near normal', and 'above normal' DIR respectively, in the ternary phase diagrams. $\times$ marks location of the reference forecast (a) $\mathbf{q} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and (b) $\mathbf{q} = (65/100, 12/100, 23/100)$.

Figure 6.7 presents probabilistic forecast maps for DIR, FMA season 2008 and 2009 for both the GLMM (Fig. 6.7.1a and 2a) and the ARM (Fig. 6.7.1b and 2b). Maps of the observed DIR category for each microregion are shown for comparison (Fig. 6.7.1c and 2c). For the FMA season 2008, the GLMM would have correctly forecast high DIR for Rio de Janeiro and microregions along the east coast and in the west of the region (see Fig. 6.7.1a). The GLMM also correctly forecast low DIR in the south. Conversely, the ARM forecasts low DIR or the reference forecast across nearly the entire region. This

model would not have provoked any epidemic warnings across the South East for FMA 2008, when a serious epidemic occurred. For FMA 2009, a forecast from the GLMM may have resulted in a false alarm for Rio de Janeiro (see Fig. 6.7.2a) but successful epidemic alerts would have been possible in areas along the east coast and in the west. Again, the ARM forecasts low DIR and the reference forecast for much of the region, but epidemic alerts for one or two microregions may have been possible using this model.

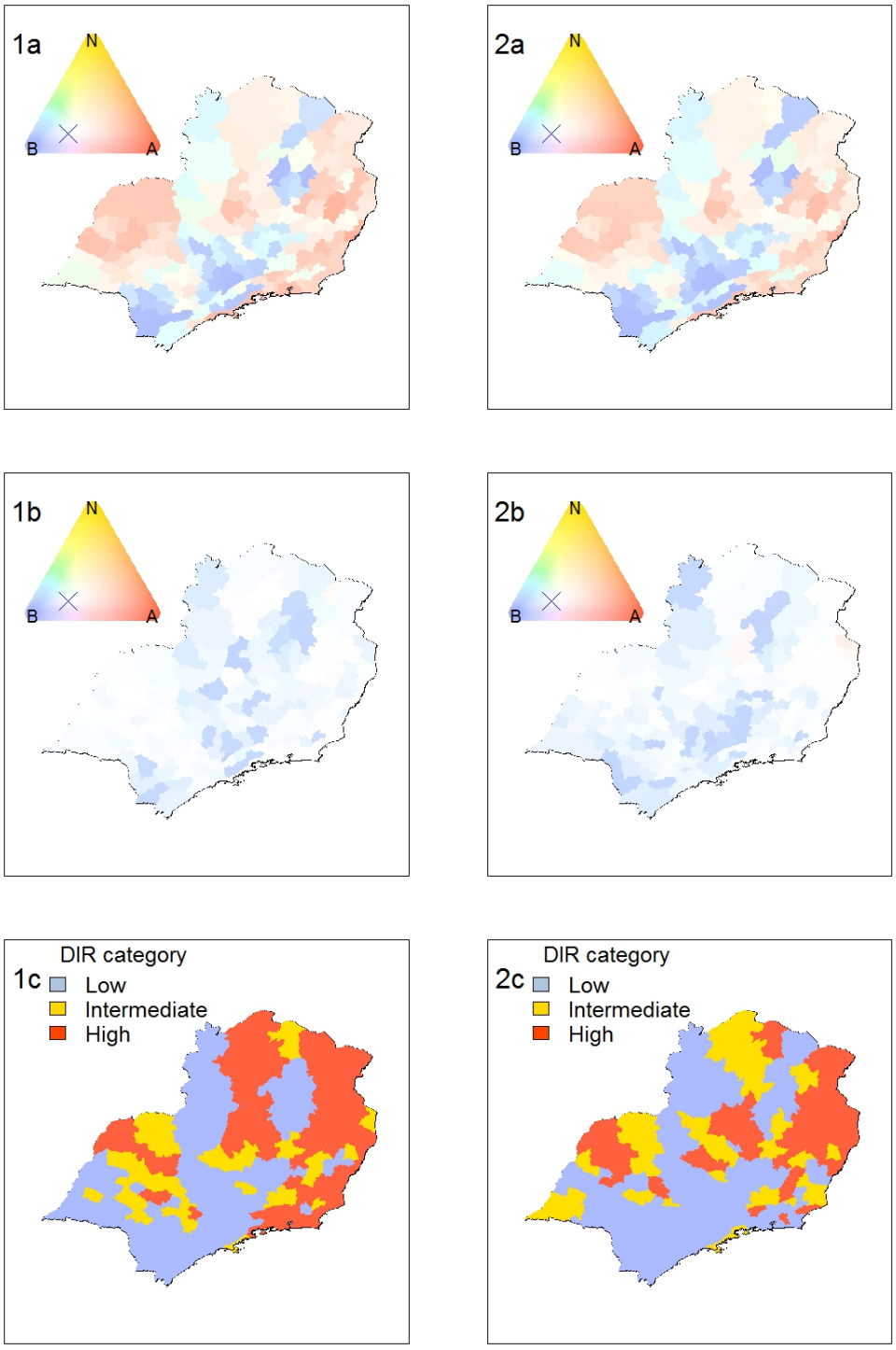


Figure 6.7: Probabilistic forecast using (a) GLMM and (b) ARM for FMA 2008 (column 1) and 2009 (column 2). (c) Corresponding observed categories. Category boundaries defined as 100 and 300 cases per 100,000 inhabitants.

So far, this investigation has revealed that the GLMM performs better than current practice for South East Brazil. Although there is a tendency for over-prediction of DIR using the GLMM (i.e. the forecasting system is more *liberal*), it is better able to capture the spatial and temporal behaviour of dengue than the ARM, which relies on dengue in the previous months. It is also capable of detecting elevated levels of DIR which is important for an early warning system to help direct the allocation of resources to cope with area-specific dengue epidemics. In order to verify how well the GLMM and the ARM predict dengue epidemics, the skill of these forecasting systems in predicting dengue exceeding an epidemic threshold, across the South East region, will next be evaluated and compared.

6.5 Evaluation of dengue forecasting systems

Probabilistic forecasts of any event can be easily evaluated by considering the set of deterministic binary forecasts obtained by choosing a range of probability decision thresholds (Mason, 1979). The GLMM and ARM can be used to predict the probability of dengue exceeding a pre-defined epidemic threshold in each microregion. As the posterior predictive distribution can be obtained for each microregion (rather than a point estimate), the probability of exceeding an epidemic threshold can be calculated. The decision to trigger an alert can be based on the probability of exceeding the threshold being greater than a specified alert level, (e.g. probability of exceedance $\geq 50\%$). Many epidemic detection algorithms have been investigated to detect epidemics (Cullen et al., 1984; Hay et al., 2002; Teklehaimanot et al., 2004). As an example, the event of dengue incidence exceeding 300 cases per 100,000 inhabitants (high incidence threshold defined by the National Dengue Control Programme (PNCD) in Brazil) will be considered. The ability of the GLMM to predict dengue epidemics across South East Brazil during the FMA season in 2008 and 2009 can be assessed using a contingency table (see Table 6.2, Jolliffe and Stephenson, 2003).

There are two ways for the forecast to be correct (either a hit or a correct rejection) and two ways for the forecast to be incorrect (either a false alarm or a miss). Cell count a is the number of events correctly forecast to occur, i.e. the number of hits; cell count b is the number of events incorrectly forecast to occur, i.e. the number of false alarms; cell

Table 6.2: The four possible outcomes for categorical forecasts of a binary event.

		Event Observed		
		Yes	No	Total
Warning	Yes	a=hits	b=false alarms	$a+b$
Issued	No	c=misses	d=correct rejections	$c+d$
	Total	$a+c$	$b+d$	$a+b+c+d=n$

count c is the number events incorrectly forecast not to occur, i.e. the number of misses; and cell count d is the number of event correctly forecast not to occur, i.e. the number of correct rejections.

The *proportion correct*, and conditional probabilities such as the *hit rate* and the *false alarm rate*, can be calculated to assess the correspondence between forecasts and observations. Observed DIR for the 3-month season FMA 2008 was compared with model predictions where the probability of an epidemic exceeded probability decision thresholds varying between 0-100%. During this season, 54 of the 160 microregions in South East Brazil experienced an 'epidemic'. The contingency table provides information on the overall predictive skill of the warning system given a specific threshold. For example, given a probability decision threshold of 50%, the proportion correct (PC), defined as the proportion of the 160 microregions for which the prediction correctly anticipated the subsequent epidemic or non-epidemic, $(a+d)/(a+b+c+d)$, was 78%. The hit rate (HR); the proportion of epidemics that were correctly predicted ($a/(a+c)$, also known as sensitivity), was 57%. Conversely, the false alarm rate (FAR); the proportion of epidemics that were predicted but did not occur ($b/(b+d)$, also known as 1-specificity), was 12% (see Table 6.3). When the probability decision threshold was lowered to 30%, PC= 79%, HR= 94% and FAR= 29%. By lowering the probability decision threshold, the hit rate for the region increases but so does the false alarm rate.

For the following FMA season in 2009, 37 of the 160 microregions in South East Brazil experienced an epidemic (DIR > 300). The system did not perform as well for this year. For a probability decision threshold of 30%, the hit rate lowered to 84% while the false alarm rate increased to 36% (see Table 6.4).

Clearly, a single set of binary forecasts does not provide a satisfactory basis for assess-

Table 6.3: Summary of contingency table results for observed DIR exceeding epidemic threshold of 300 cases per 100,000 inhabitants at varying probability decision thresholds (50%, 40%, 30%) for the 160 microregions FMA 2008 using GLMM.

Threshold	a	b	c	d	PC	HR	FAR
50%	31	13	23	93	78%	57%	12%
40%	41	24	13	82	77%	76%	23%
30%	51	31	3	75	79%	94%	29%

Table 6.4: Summary of contingency table results for observed DIR exceeding epidemic threshold of 300 cases per 100,000 inhabitants at varying probability decision thresholds (50%, 40%, 30%) for the 160 microregions FMA 2009 using GLMM.

Threshold	a	b	c	d	PC	HR	FAR
50%	16	19	21	104	75%	43%	15%
40%	21	32	16	91	70%	57%	26%
30%	31	44	6	79	63%	84%	36%

ment of the quality of the forecasting system (Mason, 2003). This is because it shows the performance of the system at only a single probability decision threshold. A complete description of predictive skill requires verification over the full range of possible thresholds. An analysis tool that accomplishes this is the Relative (or Receiver) Operating Characteristic (ROC). The ROC is a graph of the hit rate against the false alarm rate (or sensitivity against 1-specificity) for different decision thresholds (see Fig 6.8). ROC graphs have long been used in signal detection theory to depict the trade-off between hit rates and false alarm rates of classifiers (see Fawcett, 2006 and references therein).

The location of the whole curve in the unit square is determined by the intrinsic discrimination capacity of the forecasting system and the location of specific points on a curve is fixed by the probability decision threshold at which the system is operating (Mason, 2003). As the probability decision threshold varies from high to low (moving from left to right) HR and FAR vary together to trace out the ROC curve. Perfect discrimination is represented by the point (0,1) where HR= 100% and FAR= 0%. The diagonal HR=FAR represents zero skill, i.e. the forecasting system performs as well as random guessing. The area under the modelled ROC curve, abbreviated AUC (Fawcett, 2006), is a widely

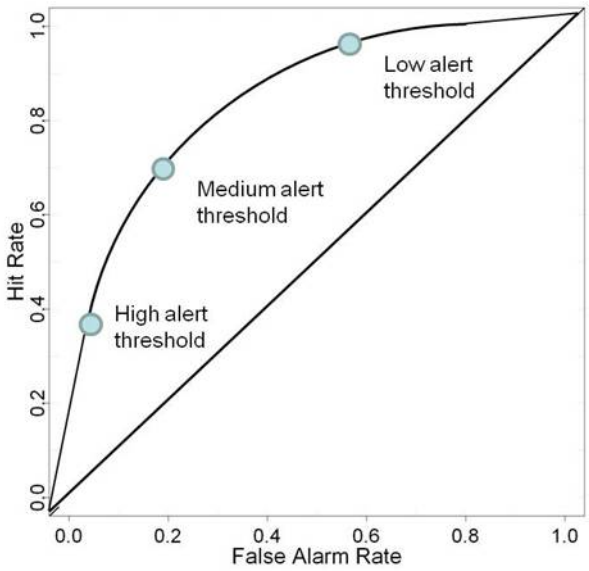


Figure 6.8: ROC curve schematic for a binary event with different probabilistic decision thresholds.

used ROC-based measure of skill. AUC characterises the quality of a forecast system by describing the system's ability to anticipate correctly the occurrence or non-occurrence of pre-defined events (Mason and Graham, 2002). The possible range of AUC is $[0, 1]$. Zero skill is indicated by AUC=0.5, i.e. area under the diagonal HR=FAR. For perfect skill, AUC=1. To test the null hypothesis that the area under the ROC curve is 0.5, i.e. the forecast has no skill, a p-value can be calculated using a Mann-Whitney U-test (see Mason and Graham, 2002). Figure 6.9 shows ROC curves for dengue epidemics for the FMA season 2008 and 2009 for the 160 microregions in South East Brazil. In Figure 6.9, AUC=0.85 (p-value $<<$ 0.05) for the FMA seaon 2008 and AUC=0.80 (p-value $<<$ 0.05) for the FMA seaon 2009 (see Fig. 6.9). This indicates that the forecasting system is significantly more skillful than randomly guessing and that the system performed better in FMA 2008.

Figure 6.10 shows the posterior predictive distributions for the FMA season 2008 and 2009 for the microregions Três Marias and Belo Horizonte, Baía de Ilha Grande, Rio de Janeiro and São Jose dos Campos. Using the GLMM with the given epidemic threshold of 300 cases per 100,000 inhabitants and a probability decision threshold of 30%, a successful epidemic alert would have been issued for Três Marias and Belo Horizonte in FMA 2008 and 2009, resulting in a hit (see Fig. 6.10.1a, $p(\text{DIR}) > 300 = 0.52$, Fig. 6.10.2a, $p(\text{DIR}) >$

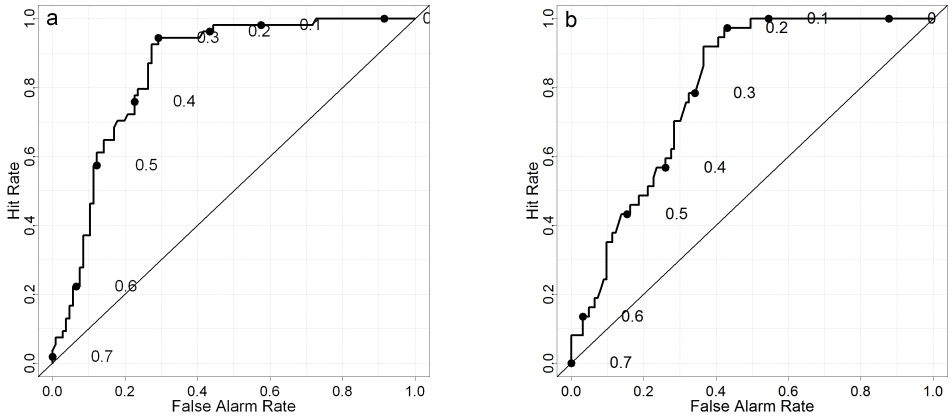


Figure 6.9: ROC curve for binary event of observed DIR exceeding the epidemic threshold of 300 cases per 100,000 inhabitants for FMA (a) 2008 and (b) 2009 using GLMM.

300 = 0.54, Fig. 6.10.1b, $p(\text{DIR}) > 300 = 0.31$ and Fig. 6.10.2b, $p(\text{DIR}) > 300 = 0.30$). In both Baía de Ilha Grande and Rio de Janeiro, the forecast would have resulted in a hit in FMA 2008, but a false alarm in 2009 (see Fig. 6.10.1c, $p(\text{DIR}) > 300 = 0.66$, Fig. 6.10.2c, $p(\text{DIR}) > 300 = 0.63$, Fig. 6.10.1d, $p(\text{DIR}) > 300 = 0.64$ and Fig. 6.10.2d, $p(\text{DIR}) > 300 = 0.57$). In São Jose dos Campos, the forecasts for both years would have resulted in correct rejections (see Fig. 6.10.1e, $p(\text{DIR}) > 300 = 0.10$ and Fig. 6.10.2e, $p(\text{DIR}) > 300 = 0.04$).

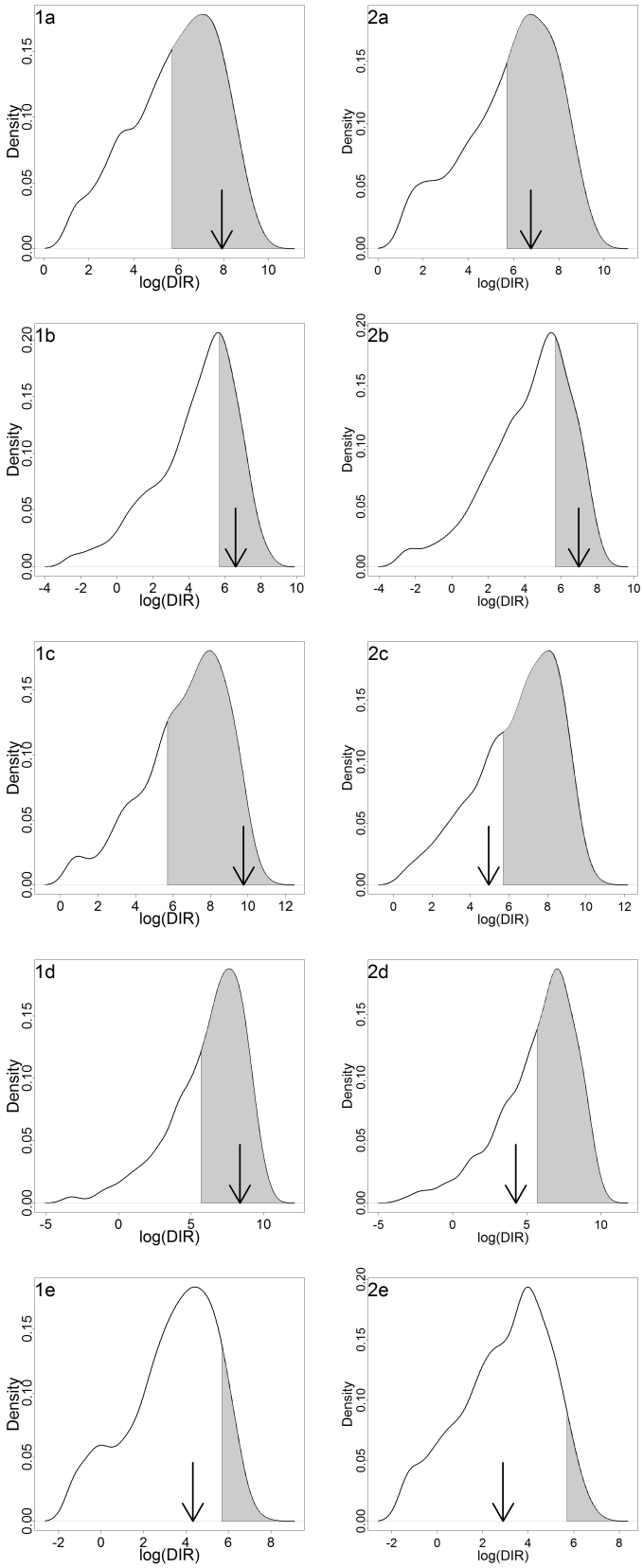


Figure 6.10: Posterior predictive distributions and probability of exceeding the pre-defined epidemic threshold of 300 cases per 100,000 inhabitants (shaded area) for FMA 2008 (column 1) and FMA 2009 (column 2) for selected microregions: (a) Três Marias, (b) Belo Horizonte, (c) Baía de Ilha Grande, (d) Rio de Janeiro and (e) São Jose dos Campos. Arrow indicates observed DIR.

For comparison, the predictive ability of the current practice ARM was evaluated. This model proved to be a lot more *conservative* than the GLMM. Figure 6.11 shows ROC curves for dengue epidemics for the FMA season 2008 and 2009 for the 160 microregions in South East Brazil using the ARM. Given a probability threshold greater than 30%, no false alarms or hits would have been issued for the FMA season, 2008. For a probability threshold of 20%, HR=59% and FAR=8%. In Figure 6.11a, AUC=0.83 (p-value << 0.05) for the FMA season 2008. This is slightly lower than the AUC for the GLMM. However, using this model for FMA season 2009, a probability threshold of 20% would give HR=70% and FAR=5% and AUC=0.95 (p-value << 0.05, see Fig. 6.11b).

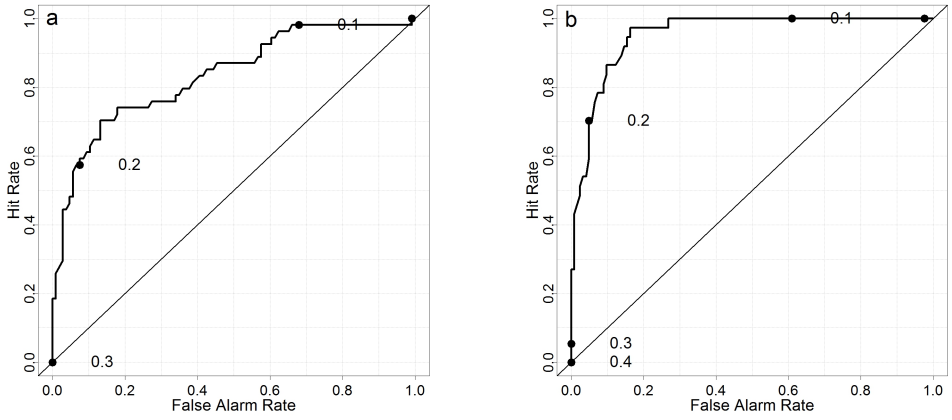


Figure 6.11: ROC curve for binary event of observed DIR exceeding the epidemic threshold of 300 cases per 100,000 inhabitants for FMA (a) 2008 and (b) 2009 using ARM.

Although the ARM is a much more conservative model than the GLMM, it appears that some information could be gained by including dengue relative risk observed 3 months previously in the model framework. In the following section the GLMM and ARM will be combined into a single dengue prediction model and the performance of this model will be evaluated.

6.6 Combining GLMM with current practice

Although climate variables capture some of the inter-annual variability in dengue incidence, there are clearly unobserved confounding temporal factors missing from the model

which may account for the epidemic cycle present in the study period. The number of dengue cases observed several months previously might indicate the presence of increased mosquito populations or the circulation of a new dengue serotype to which the human population is not immune. A lagged dengue relative risk term could then act as a surrogate for unobserved and unmeasured spatio-temporal confounding factors in the model. The GLMM and ARM were combined by including the 3 month lag dengue relative risk term in the GLMM model formulation (Eqn. 5.6) and fitted to the data April 2001 - December 2009. The model specification for this 'combined' GLMM is as follows:

$$
\begin{aligned}
&y_{st}|\mu_{st} \sim \text{NegBin}(\mu_{st}, \kappa) \\
&\log(\mu_{st}) = \log(e_{st}) + \alpha + \delta_{t'(t)} + \gamma_0 \log(\frac{y_{st-3}}{e_{st-3}}) + \sum_j \gamma_j w_{jst} + \sum_j \beta_j x_{jst} + \phi_s + \upsilon_s \\
&\kappa \sim \text{Ga}(0.5, 0.0005) \\
&\alpha \sim \text{U}(-\infty, +\infty) \\
&\phi_s \sim \text{N}(0, \sigma_\phi^2) \\
&\upsilon_s \sim \text{CAR}(\sigma_\upsilon^2) \\
&\tau_\phi \sim \text{Ga}(0.5, 0.0005) \\
&\tau_\upsilon \sim \text{Ga}(0.5, 0.0005),
\end{aligned}
$$

with independent diffuse Gaussian priors (mean 0, precision 1×10^{-6}) taken for the fixed effects β_j $(j = 1, \ldots, 3)$, γ_j $(j = 0, 1, 2)$ and $\delta_{t'(t)}$ with $t'(t) = 2, \ldots, 12$ (see Chapter 5, Section 5.5 for more details). It is interesting to see how the inclusion of this term in the GLMM formulation will affect the parameter estimate for the climate covariates; the other source of spatio-temporal information in the model. Table 6.5 shows the parameter estimates and credible intervals for climate covariates, lagged SMR and overdispersion parameter obtained from the GLMM, the ARM and their combination in the combined GLMM. The addition of the lagged SMR term does not alter the distribution of the climate estimates by very much. This indicates that the climate covariates in the model are robust predictors and further supports the argument to include the ENSO index in the model. The parameter estimate for the climate and lagged SMR covariates slightly reduce in magnitude when included together in the combined GLMM. This is because these covariates account for some of the variation explained by other factors missing from the model formulation when fitted in separate models. For example, climate variation is missing in the ARM. By including the lagged SMR in the combined GLMM, the DIC

and overdispersion parameter slightly decrease (see Table 6.5).

Table 6.5: Parameter estimates and 95% credible intervals for climate covariates, lagged SMR and overdispersion parameter from GLMM, ARM and their combination.

	GLMM	ARM	Combined GLMM
precipitation	0.314 (0.247, 0.376)	-	0.305 (0.246, 0.365)
temperature	0.561 (0.489, 0.633)	-	0.491 (0.424, 0.555)
ONI	-0.472 (-0.520, -0.426)	-	-0.410 (-0.454, -0.366)
SMR lag 3	-	0.253 (0.245, 0.260)	0.215 (0.206, 0.223)
κ^{-1}	2.531 (2.468, 2.597)	5.565 (5.444, 5.698)	2.126 (2.071, 2.182)
DIC	92022.7	102857	89869.3

Figure 6.12 shows the DIR for the South East region using the GLMM, ARM and combined GLMM from April 2001 - December 2009. By including past dengue cases in the model, timely epidemic warnings would have been possible in both 2002 and 2008 for the region. The addition of lagged dengue into the combined GLMM (see Fig. 6.12c) better captures the epidemic cycle during the time period compared to the GLMM (see Fig. 6.12a) and provides timely estimates of changes in DIR compared to the ARM (see Fig. 6.12b).

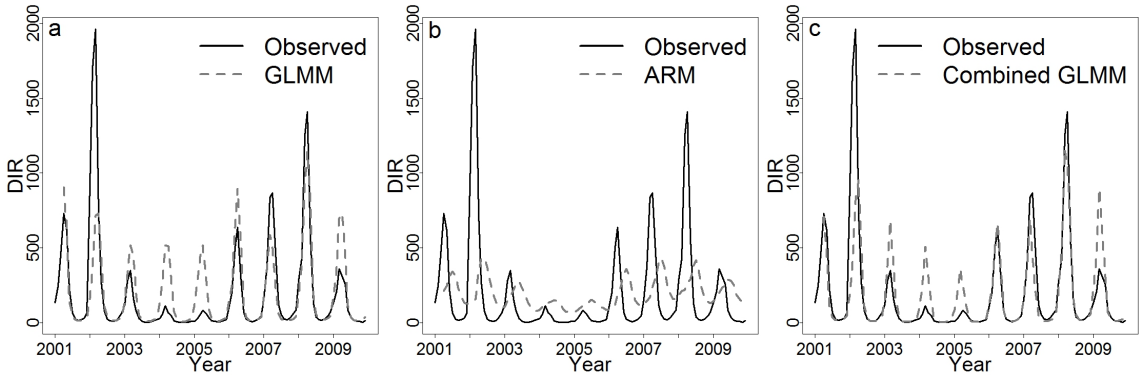


Figure 6.12: Total observed (solid line) and model fit (dashed line) DIR from April 2001 - December 2009 for (a) GLMM, (b) ARM and (c) combined GLMM.

To assess the predictive ability of the combined GLMM, this model was refitted to the first 7 years of data (April 2001 - December 2007) and posterior predictive distributions of DIR at each microregion was obtained for January 2008 - December 2009 (see Section 6.3). Figure 6.13 shows probabilistic forecast maps for DIR, FMA season 2008 and 2009 using

the combined GLMM. Compared to the probabilistic forecast given by the GLMM for FMA season 2008 (see Fig. 6.7.1a), there was more information gain (i.e. compared to the reference forecast) in the forecast from the combined GLMM, particularly for DIR in the high risk category. For example, along the east coast and central north of the region the combined GLMM issued a high probability of DIR exceeding 300 cases per 100,000 inhabitants (darker shades of red). However, this model could have also resulted in several false alarms being issued (e.g. south west of the region, see Fig. 6.13.1a). For the FMA season 2009, the combined GLMM correctly predicted high risk for some microregions in the centre of the region and on the east coast that were previously predicted with less certainty using the GLMM (see Fig. 6.13.2a).

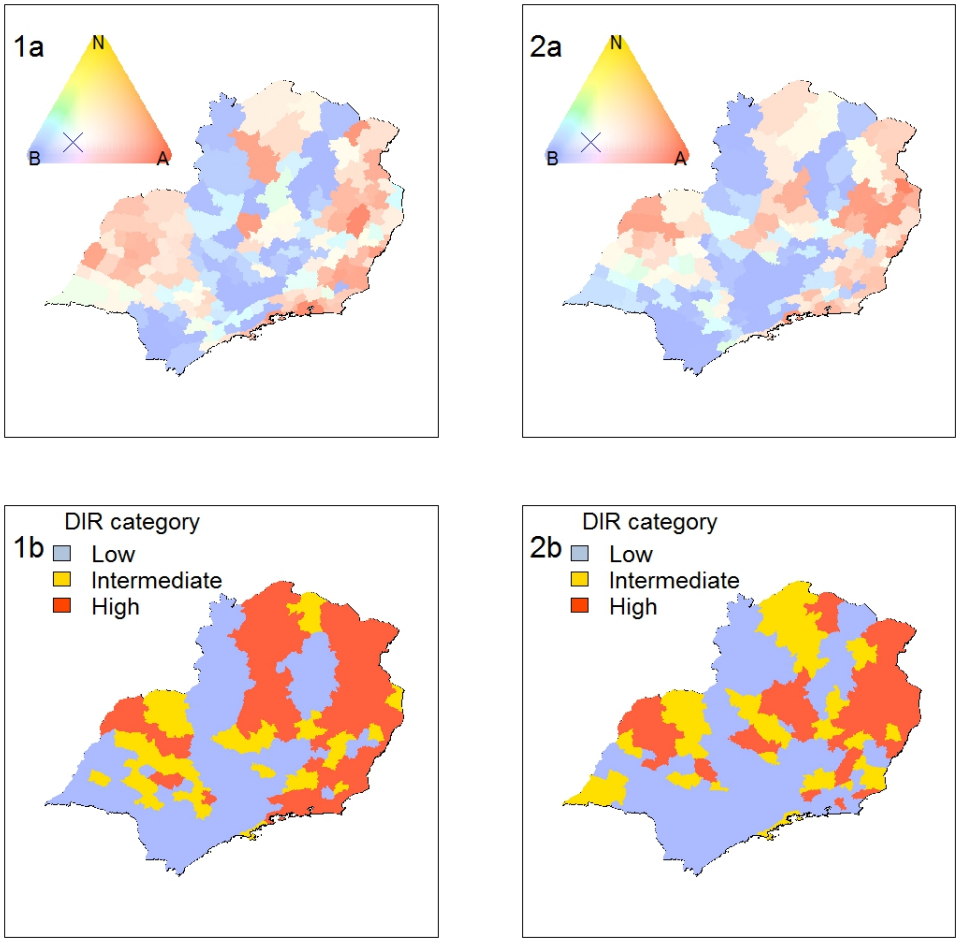


Figure 6.13: (a) Probabilistic forecast using combined GLMM. (b) Corresponding observed categories for FMA 2008 (column 1) and 2009 (column 2). Category boundaries defined as 100 and 300 cases per 100,000 inhabitants.

Using the combined GLMM, the binary classification of dengue incidence exceeding 300 cases per 100,000 inhabitants was evaluated with contingency tables and ROC analysis. Table 6.6 shows contingency table results for FMA season 2008. Compared to the GLMM (see Table 6.3), the hit rate increased given a probability decision threshold of 50% and the false alarm rate decreased slightly. At a probability decision threshold of 30%, both the hit rate and false alarm rate decrease slightly. However, for FMA 2009, using the combined GLMM results in a much greater hit rate and a lower false alarm rate at the three given probability thresholds (see Table 6.7) compared to the GLMM (see Table 6.4).

Table 6.6: Summary of contingency table results for observed DIR exceeding epidemic threshold of 300 cases per 100,000 inhabitants at varying probability decision thresholds $(50\%, 40\%, 30\%)$ for the 160 microregions FMA 2008 using combined GLMM.

Threshold	a	b	c	d	PC	HR	FAR
50%	34	10	20	96	81%	63%	9%
40%	40	18	14	88	80%	71%	17%
30%	47	24	7	82	81%	87%	23%

Table 6.7: Summary of contingency table results for observed DIR exceeding epidemic threshold of 300 cases per 100,000 inhabitants at varying probability decision thresholds $(50\%, 40\%, 30\%)$ for the 160 microregions FMA 2009 using combined GLMM.

Threshold	a	b	c	d	PC	HR	FAR
50%	25	10	12	113	86%	68%	8%
40%	31	16	6	107	86%	84%	13%
30%	34	23	3	100	84%	92%	19%

Figure 6.14 shows ROC curves for dengue epidemics for the FMA season 2008 and 2009 for the 160 microregions in South East Brazil using the combined GLMM. In Figure 6.14a, the area beneath the ROC curve increased to AUC=0.88 (p-value $<<$ 0.05) for the FMA season 2008. This area is greater than the area given by both the GLMM and the ARM (see Table 6.1). Using this model for FMA season 2009, AUC=0.94 (p-value $<<$ 0.05, see Fig. 6.9) which is an improvement compared to the GLMM, but slightly lower than for the ARM in the season.

In general, the combined GLMM appears to provide a small but significant enhance-

Table 6.8: Area under the modelled ROC curve (AUC) for GLMM, ARM and combined GLMM for epidemic prediction of FMA season 2008 and 2009.

	2008	2009
GLMM	0.85	0.80
ARM	0.83	0.95
Combined GLMM	0.88	0.94

ment to dengue incidence prediction compared to the GLMM developed in this thesis. Although the combined GLMM still produces a considerable number of false alarms compared to the ARM, a model that is able to correctly predict geographically specific increased dengue incidence the majority of the time might be desirable to the Brazilian Ministry of Health, despite having to endure potential false alarms.

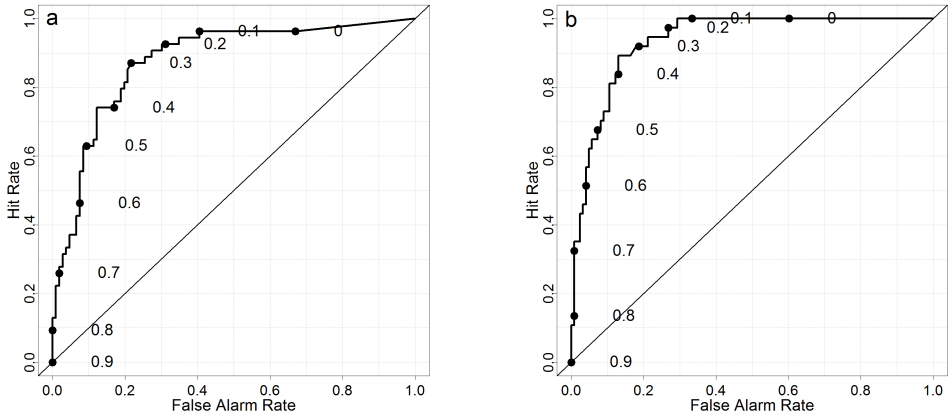


Figure 6.14: ROC curve for binary event of observed DIR exceeding the epidemic threshold of 300 cases per 100,000 inhabitants for FMA (a) 2008 and (b) 2009 using combined GLMM.

6.7 Conclusion

In this chapter, a comparison was made between the GLMM developed in this thesis and a simple mathematical model of current practice. The GLMM was found to be a more adequate model with a lower DIC. The GLMM also better captured the overall spatial and seasonal distribution of dengue compared to the ARM, although the GLMM

was more prone to over-estimating dengue incidence rates. Predictions from both models were evaluated by obtaining posterior predictive distributions from the GLMM and the ARM, fitted to dengue cases in the South East of Brazil April 2001 - December 2007 and comparing predictions to out-of-sample data, January 2008 - December 2009. The comparison benefited from a novel procedure for visualising ternary probability forecasts. The method conveys more information from a probability forecast than traditional methods and provides an understanding of the information gain of the forecast compared to a meaningful reference forecast. The procedure allows users to quickly identify regions where the forecasting system is more certain that the predicted outcome will occur. In both 2008 and 2009 the GLMM was able to predict high dengue incidence with some certainty while the ARM model predicted low DIR or the reference forecast for most of the region. The ARM would have rarely provoked an epidemic warning.

The results demonstrate that the GLMM performs better in statistical terms than the current practice model, based on measures of model adequacy and time series of the posterior predictive distributions. However, to ascertain if the developed GLMM could provide an improved prediction tool for dengue epidemics in South East Brazil, both models were verified using ROC analysis to assess the spatial capability of the forecasting systems for two out-of-sample seasons, FMA 2008 and 2009. As posterior predictive distributions are obtained for each microregion and time period, the probability of DIR exceeding a pre-defined epidemic of 300 cases per 100,000 inhabitants was determined. For the FMA season, 2008, the forecasting system demonstrated considerable skill, compared to a system that issues the reference forecast for each season/year. When the probabilistic forecast information is summarised into a binary classification of epidemic or non-epidemic, the current practice model performs nearly as well as the GLMM for FMA 2008, based on the AUC. The skill of the GLMM forecasting system reduced slightly in 2009 but was still a better system than random guessing. However, the skill for the current practice model was better in 2009.

Results demonstrated that there may be some benefit in combining the prediction models. The inclusion of lagged dengue risk in the model may act as a surrogate for increased mosquito populations or serotype introduction. When this term was included in the model formulation, the parameter estimates for the climate information did not change much, further indicating that the climate covariates are robust predictors of dengue relative risk. By adding an extra source of spatio-temporal information in the GLMM,

the dengue predictions improved in both space and time. If the model developed in this thesis were to be implemented operationally, it is recommended that for prediction purposes, the Brazilian Ministry of Health also consider past dengue relative risk in the model formulation and adopt the combined modelling approach.

By lowering the probability decision threshold, the hit rate increases but so does the false alarm rate. Optimal probability decision thresholds are sometimes determined as the point where the ROC curve intersects the negative 45° line (where sensitivity=specificity or HR=1-FAR) or the point where the distance from the HR=FAR line is greatest (Pepe, 2004). In practice, the choice of epidemic threshold and probability decision thresholds should be selected by decision makers based on expert opinion and available resources.

Communicating information contained within a probabilistic forecast presents a challenge. It is hoped that the visualisation method presented here facilitates the interpretation of the probabilistic forecasts of dengue incidence rates for each microregion across South East Brazil. The proposed method can be extended by incorporating information about the past skill of the forecasting system to the map by plotting each forecast as a circle whose radius is proportional to some measure of skill (Jupp et al., 2010). However, the time period for which dengue data is available is currently too short to calculate a reliable reference forecast and subsequent skill scores. ROC graphs are useful tools for visualising and evaluating forecasting systems. For evaluating an epidemic prediction system for a given time period in space, a binary classification of a disease exceeding an epidemic threshold is useful for decisions related to disease epidemic warnings and interventions. However, extension of this analysis technique from binary classification problems to multi-class problems provides a much wider applicability of this technique (e.g. Everson and Fieldsend, 2006). Combining such techniques with the proposed visualisation method may prove useful for evaluating climate forecasting systems where droughts and cold snaps are as important as floods and heat waves. The evaluation of the forecasting system across a range of categories is then desirable.

Chapter 7

Conclusions

7.1 Summary of main findings

The main contribution of this research is the development of a modelling framework to predict spatio-temporal variations in dengue risk. The model combines climatic and non-climatic factors in the model parameterization to correctly quantify variability captured by climate information. This is the first study that includes climate information to model spatio-temporal variations in dengue incidence in Brazil, at the national and regional level. An obstacle in modelling such a complex disease in space and time is the presence of extra-Poisson variation. This was primarily modelled using a negative binomial generalised linear model that includes an additional scale parameter to account for such overdispersion. Although the generalised linear model accounted for extra variation by the inclusion of climate and non-climate variables and interactions between the annual cycle and geographic zones, there was still a large proportion of the variance that was unexplained. Therefore, before inference could be made as to the sensitivity of dengue risk to climate variability, a generalised linear mixed model was adopted which included spatially unstructured and structured random effects in the linear predictor. The modelling framework was extended for South East Brazil; a region where the GLM performed best and there are a large number of densely populated urban centres, which could benefit from a climate informed dengue early warning system. Spatially unstructured random effects explicitly model microregion specific extra-Poisson variation in addition to the scale parameter in the negative binomial GLMM, while spatially structured random effects

allow for correlated heterogeneity between microregions.

The spatio-temporal hierarchical model was fitted using a Bayesian estimation framework. Posterior predictive distributions for disease risk were derived at each spatial location for a given month or season. This allowed probabilistic forecasts to be issued. A thorough evaluation of the forecast skill of dengue epidemic warnings using out-of-sample data was conducted. The spatio-temporal hierarchical model was compared to a simple conceptual model of current practice, based on dengue cases three months previously. It was found that a model based on past dengue cases alone was of little use for an epidemic early warning system. However, incorporating this information into the developed spatio-temporal hierarchical model enhanced dengue predictions in time and space. In the event of implementing such a forecasting system operationally, it is recommended that the Brazilian Ministry of Health also consider past dengue relative risk in the model formulation and adopt the combined modelling approach.

The analysis benefited from a novel procedure for visualising ternary probability forecasts. The method conveys more information from a probability forecast than conventional methods and provides an understanding of the information gain of the forecast compared to climatology. This is a pioneering example of visually conveying probabilistic ternary forecasts in epidemiology, based on techniques adopted from climate science.

A major obstacle to developing a climate-driven dengue model is the lack of climate and disease data over long time periods. The model parameterisation would benefit from the inclusion of more epidemic years to address these problems. Other limitations include the lack of serological and entomological data. This meant that it was difficult to empirically determine if the ENSO index was spuriously correlated with dengue transmission. However, ENSO remained a robust predictor after the inclusion of past dengue cases in the model, which may act as a surrogate for missing information such as increased mosquito populations or serotype introduction. The primary biological basis for the ENSO-dengue relationship is that ENSO drives local variations in climate, and local climate variations affect dengue transmission. The model indicated that precipitation and temperature are positively associated with dengue relative risk. However, the negative association of dengue with ONI may be indicative of optimum local climate conditions for the proliferation of dengue-carrying mosquito. Intra-seasonal changes during La Niña summers have been observed across areas of South East Brazil (Grimm, 2004). Therefore, a finer tem-

poral resolution than seasonally averaged climate variables may be required to capture such reversals and their impact on dengue transmission. As these features are not easily explained by changes in the remote influences caused by La Niña events, data concerning SST anomalies in the Atlantic Ocean may prove useful in determining local climate conditions in South East Brazil in advance (Jose Marengo, pers. comms). Although the robustness of the ENSO index for predicting dengue needs to be verified as more data become available, the effects of ENSO should be taken into account in future epidemic forecasting for public health preparedness.

Another potentially important component missing from the model is the seasonal movement of human hosts around Brazil. The proximity matrix used to formulate the CAR prior (see Chapter 5, Section 5.4.2) for the spatially structured random effects in the GLMM, assumes a simple local structure where each microregion is dependent only on its neighbours. However, certain areas may be more closely related, in terms of dengue transmission, to remote areas connected by air or road transport links, rather than neighbouring microregions. IBGE have released a new study entitled 'Areas of Influence of Cities'[1] based on research into the Brazilian urban network. A hierarchy of urban centres in defined based on the flow of good and services, including air and road travel. A proximity matrix based on this hierarchical matrix might improve the correlation structure within the model. As epidemics tend to start in Rio de Janeiro and spread to other parts of the country (Teixeira et al., 2005), information relating to the movement of human hosts may be of even greater importance outside of the South East region.

7.2 Applying model framework to other regions in Brazil

Although this thesis has focused primarily on developing a model for South East Brazil, it is interesting to question how well the modelling framework developed in Chapter 4 for Brazil as a whole, could explain dengue relative risk in other Brazilian regions. Preliminary analyses of extending the modelling framework for the North East region have begun. Due to the varied climate and geography in the North East region (Amazon Rainforest, Caatinga, Cerrado, North East Atlantic Rainforest), interactions between zone and climate variables, explored in Chapter 4, are important for this region. Fig-

[1] http://www.ibge.gov.br/english/geociencias/geografia/regic.shtm, [accessed 5 September 2010]

ure 7.1 shows the observed and forecast DIR for the March-May (MAM, the peak dengue season for this region) in 2002. This model is able to correctly predict very high DIR in certain places. Although non-linear terms were not found to be important in Chapters 4 and 5, inspection of residuals plots for the North East Brazil model indicate that there is stronger evidence to suggest that non-linear climatic terms such as precipitation squared could be important in this region. Therefore, the possibility of non-linear relationship between dengue and climate will be explored further in future model development for the North East and remaining Brazilian regions.

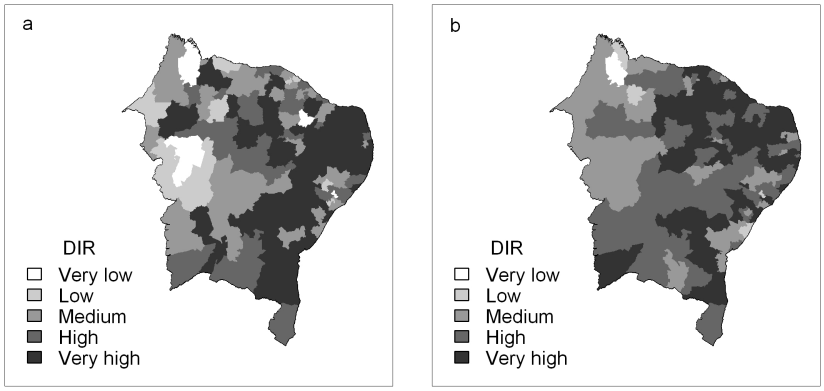


Figure 7.1: (a) Observed and (b) model fit DIR using GLMM for North East Brazil, MAM season 2002.

7.3 Extending lead-time by using climate forecasts

An issue that this thesis has not considered is to what extent the lead time of dengue risk prediction could be extended using seasonal climate forecasts. In practice, observed climate could be replaced by climate forecasts which might extend the lead time beyond that offered by using lagged observations. The next stage of this research would be to assess the predictive validity of the model when replacing 'observed' with 'hindcast' (i.e. retrospective forecasts made for a historical period in pseudo-operational mode) climate variables. 'Hindcast' climate data are available from forecasting systems such as the UK Met Office seasonal forecasting system (Graham et al., 2005) and the European Centre for Medium Range Forecasts (ECMWF) System 3 (Anderson et al., 2007). These systems typically produce ensemble predictions with lead times up to 6 months. A

preliminary analysis using seasonal climate forecasts from the Met Office GloSea3 model in a statistical dengue model was presented in Lowe et al. (2009).

By replacing 'observed' with 'hindcast' climate variables, a dengue prediction could be made several months ahead of the dengue season of interest. For example, to predict dengue incidence for March 2011, the model could be run in November 2010 using the observed ONI for August-October 2010 (6 month lag), and precipitation and temperature forecasts for DJF 2010-2011 issued in November 2010 (see Fig 7.2). This would provide a four month lead time, which could allow time for the allocation of resources to interventions such as preparing health care services for increased numbers of dengue patients and educating populations to eliminate mosquito breeding sites. If past dengue cases were to be included in the model formulation, as in the combined GLMM (see Chapter 6), dengue cases four month previous could be used as a best guess for dengue three months previous (the preferred lag). This would allow a dengue prediction to be made using climate forecasts and also current dengue risk four months ahead of the peak dengue season in the combined GLMM framework. In order to issue the most accurate and up-to-date epidemic predictions, forecast climate should be replaced with observed climate as time progresses and 'past' dengue cases should be updated as they are reported, so that epidemic alerts can be re-issued to public health decision makers. A benefit of this multi-staged early warning approach is that response plans can be gradually modified as dengue forecast certainty increases. This would give public health decision makers several opportunities to weigh the costs of response actions against the risk of an impending dengue epidemic.

The efficacy of a climate-based early warning system will depend on the skill of the forecasting system. One such system that is operational in Brazil is the EUROBRISA initiative (Coelho et al., 2006) which is a multi-model combined and calibrated system that produces one-month lead precipitation forecasts for the following three-month season. Figure 7.3 shows a verification skill map of the integrated EUROBRISA system for the DJF season. There appears to be a weak positive correlation between observed and forecast precipitation anomaly for this season over parts of South East Brazil. Therefore, dengue early warnings for the FMA peak dengue season in this region, using the EUROBRISA precipitation forecast for DJF, might be feasible.

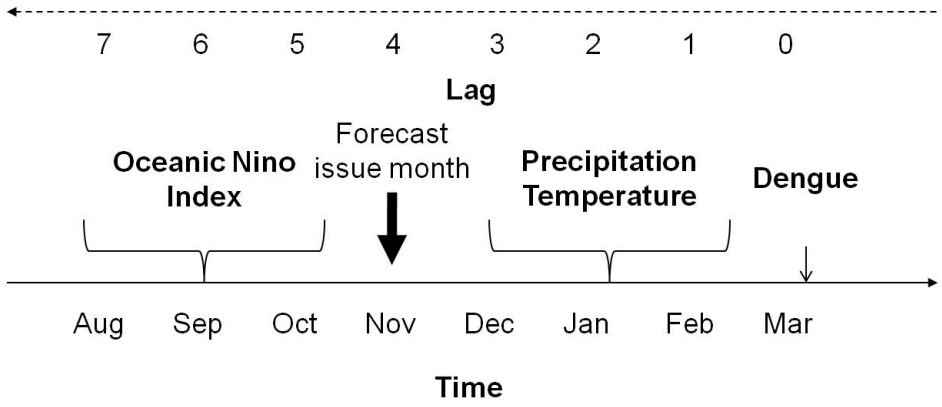


Figure 7.2: Schematic to show time lags between dengue month of interest (e.g. March), 3-month average precipitation and temperature lagged 2 months prior to dengue month (e.g. December-February) and ONI lagged 6 months prior to dengue month (e.g. August to October, 4 months prior to average precipitation and temperature). A four month lead time could be gained using a forecasting system such as EUROBRISA.

7.4 Considerations for operational early warning systems

The spatio-temporal hierarchical model is intended to become part of a newly established climate and health observatory in Brazil, operated by FIOCRUZ and INPE[2]. However, before implementing such an operational system, several technical issues need to be considered. For example, the spatial scale of the system affects the type of response activity that could be implemented. At the microregion level, interventions such as health care provisions may be possible but vector-control efforts may be more difficult to target. Probability decision thresholds should be carefully designed to minimise false alarms and false negatives (i.e. failing to predict that an epidemic will occur) and should correspond with the epidemic response capabilities in specific locations. An important issue is the consideration of future interventions in the model framework. If the Brazilian health services respond to an early warning of a dengue epidemic and take measures to reduce the impact, a false alarm may in fact be a successful intervention.

The usefulness of the model depends on the ability to obtain reliable and up-to-date

[2]http://www.inpe.br/noticias/arquivos/pdf/observatorium.pdf, [accessed 24 August 2010]

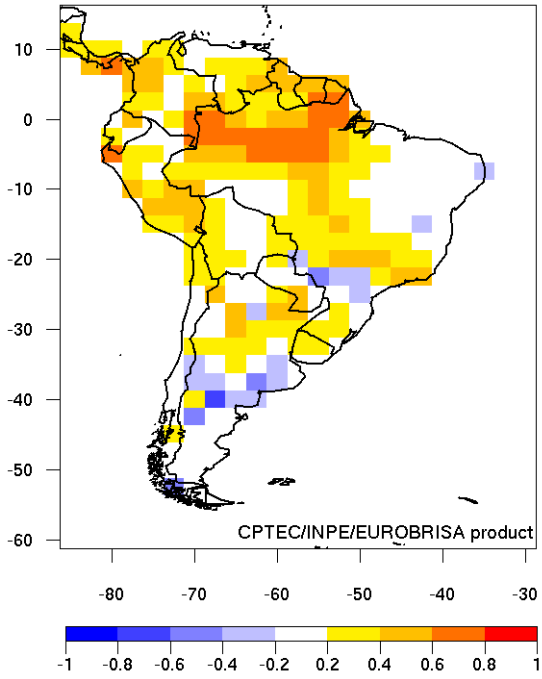


Figure 7.3: Correlation between forecast and observed precipitation anomaly using integrated EUROBRISA forecasting system for period 1981-2005. Forecasts issued in November, valid for DJF season.

information for both dengue and the factors critical to dengue incidence in time for effective responses to be implemented. Access to frequently-updated climate information is an important requirement for the development of integrated early warning systems for dengue and other climate sensitive diseases. However, time delays in obtaining real-time information for both disease cases and climate forecasts and observations could hinder the ability to provide warnings far enough in advance. As new disease/climate data becomes available over time, the early warning system should be monitored, evaluated and refitted. Spatial demographic data from the census (and interim projections) should also be updated when necessary.

Another important consideration is the dissemination and visualisation of early warnings of increased level of disease risk to public health decision makers. It is vital to train public health decision makers on how to interpret and use disease risk forecasts, including awareness about disease and climate forecast limitations to avoid misinterpretation and/or over interpretation.

All analyses in this thesis were conducted using the R statistical language. The MCMC

sampling was performed in WinBUGS, called by R using the R2WinBUGS package. These software are freely available via the Internet. This could prove useful when implementing prediction models within health surveillance systems in developing countries where investment in resources is limited.

7.5 Summary

This thesis has highlighted the potential for incorporating climate information into a spatio-temporal dengue epidemic early warning system for Brazil. Despite the limitations of the model and the difficulties involved in making such a system operational, it is hoped that this spatio-temporal dengue prediction model is a step towards the development of a useful decision making tool for the Brazilian health services. This work has strongly benefited from the inter-disciplinary collaboration between statisticians, climate scientists and public health specialists in Brazil. This model could be extended to other regions in the world where climate-sensitive infectious diseases (e.g. cholera, malaria, leptospirosis, plague) present a burden to public health infrastructure, particularly in developing countries. The developed framework will be used as a starting point to investigate whether climate information could be valuable in an early warning system for malaria in Malawi as part of the EU FP7 funded project entitled 'Quantifying Weather and Climate Impacts on Health in Developing Countries'.

The applied methodology used in this thesis exploits recent advances in spatio-temporal hierarchical mixed modelling. An advantage of implementing the model in a Bayesian framework is the ability to address specific public health issues in terms of probabilities. In view of this, it is suggested that this approach could be applied to model a wider range of spatio-temporal climate-related impacts, including agricultural, hydrological and geophysical hazards.

Appendix A

An algorithm for fitting generalised linear models

The following provides a brief description of the iterative re-weighted least squares (IRLS) algorithm. This algorithm can be used to fit generalised linear models (GLMS) by obtaining maximum likelihood estimates of the parameters $\boldsymbol{\theta}$ in the linear predictor $\eta_i = g(\mu_i)$. For further details, see McCullagh and Nelder (1989).

Given a trial estimate of the parameters $\hat{\boldsymbol{\theta}}$, an estimated linear predictor $\hat{\eta}_i = \hat{\boldsymbol{\theta}}\mathbf{x}_i'$ can be calculated and used to obtain the fitted values $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$. Using these quantities, an adjusted dependent variable

$$z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i)\frac{d\eta_i}{d\mu_i} \tag{A.1}$$

can be calculated where the derivative of the link function is evaluated at the trial estimate $\hat{\mu}$. Next, iterative weights can be defined as

$$w_i^{-1} = V(\mu_i)(\frac{d\eta_i}{d\mu_i})^2, \tag{A.2}$$

where $V(\mu_i)$ is the variance function evaluated at the trial estimate $\hat{\mu}_i$. This weight is inversely proportional to the variance of the adjusted dependent variable z_i given the current estimates of the parameters.

Finally, an improved estimate of $\boldsymbol{\theta}$ is obtained by regressing the adjusted dependent

variable z_i on the predictors $\mathbf{x}_i'$ using the weights w_i. The weighted least-squares estimate is then given by

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z},$$

where $\mathbf{X}$ is the matrix of explanatory variables, $\mathbf{W}$ is a diagonal matrix of weights with entries w_i given by the inverse of Equation A.2 and $\mathbf{z}$ is a response vector with entries z_i given by Equation A.1.

The procedure is repeated until changes in successive estimates are sufficiently small.

Appendix B

Bayesian framework and MCMC

B.1 Bayesian hierarchical modelling

The specification of a Bayesian hierarchical model involves a probability model $p(\mathbf{y}|\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_i, \ldots, \theta_p)$ and a prior distribution $p(\boldsymbol{\theta})$. In the Bayesian approach, the parameters are considered 'random quantities' rather then fixed constants. Therefore, the statistical model becomes a joint probability distribution for both the data and the parameters $p(\mathbf{y}, \boldsymbol{\theta})$ and the likelihood is the conditional distribution of $\mathbf{y}$ given the parameter values, i.e. $p(\mathbf{y}|\boldsymbol{\theta})$. Using the definition of conditional probability, $p(\mathbf{y}, \boldsymbol{\theta})$ is related to the likelihood via

$$p(\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

Here, the likelihood function $p(\mathbf{y}|\boldsymbol{\theta})$ is the product of n independent negative binomial distributions (Eqn. 4.2):

$$p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\theta_i).$$

The prior $p(\boldsymbol{\theta})$ expresses the uncertainty about $\boldsymbol{\theta}$ before taking the data into account. A hierarchical model can be created by parameterising the prior distribution with an unknown 'hyperparameter' $\boldsymbol{\vartheta}$ which has its own 'hyperprior' distribution $p(\boldsymbol{\vartheta})$. Various specifications of the prior and hyperprior distributions are possible (see Mollie, 1996; Bernardinelli et al., 2007) but they are usually chosen to be 'non-informative'. A non-informative prior has little influence on the posterior distribution.

Using Bayes Theorem, a posterior probability distribution for the parameters and hyper-parameters, given the observed data can be derived

$$p(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}). \tag{B.1}$$

The posterior $p(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\mathbf{y})$ expresses the uncertainty in $(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ given the observed data. Point estimates of $\boldsymbol{\theta}$ can be obtained from the posterior distribution. For example, one estimate of $\boldsymbol{\theta}$ is the mean of the posterior distribution:

$$\bar{\theta} = E[\boldsymbol{\theta}|\mathbf{y}] = \int_{\boldsymbol{\theta}} \int_{\vartheta} \boldsymbol{\theta} p(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\mathbf{y}) d\boldsymbol{\vartheta} d\boldsymbol{\theta}.$$

In other words, a point estimate of the set of relative risks can be obtained by calculating the posterior mean.

B.2 Estimation by Markov Chain Monte Carlo (MCMC)

Direct mathematical derivation of the posterior $p(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\mathbf{y})$ from Equation B.1 involves a high-dimensional integration to obtain the constant of proportionality (the normalising constant) and is not analytically tractable (Bailey, 2001). Therefore, simulation methods such as Monte Carlo integration are often used to approximate $p(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\mathbf{y})$. Markov Chain Monte Carlo (MCMC) methods simulate a sequence of parameter values using a Markov chain. The convergence of the values from this Markov chain to a stationary distribution, which is assumed to be the posterior distribution, must be assessed. The aim is to generate a sample from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$. A Markov chain can be constructed so that the equilibrium distribution is $p(\boldsymbol{\theta}|\mathbf{y})$. This is achieved by using algorithms such the Gibbs Sampler (German and German, 1984) which is a special case of the general framework of Metropolis et al. (1953) and Hastings (1970). When implementing MCMC, it is important to determine how long the simulations should be run and to discard a number of initial 'burn-in' iterations (Gilks et al., 1996). Sampled chains thereafter are assumed to provide approximately correct samples from the posterior distribution of interest. Saving all simulations from an MCMC run can use a large amount of storage, especially when consecutive iterations are highly correlated and a longer run is needed. It is sometimes convenient to save only every k^{th} iteration ($k > 1$) to reduce the amount of data saved from a MCMC run. This is referred to as *thinning* the chain (Raftery and Lewis, 1996). Markov chain simulation can be used to summarise a posterior distribution.

For example, the posterior expectation of any function $h(\boldsymbol{\theta})$ is defined as an integral

$$E[h(\boldsymbol{\theta})|\mathbf{y}] = \int_{\boldsymbol{\theta}} h(\boldsymbol{\theta})p(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\mathbf{y})d\boldsymbol{\theta},$$

where the integral has as many dimensions as $\boldsymbol{\theta}$. This can be approximated by the sample mean of values from the Markov chain

$$\frac{1}{G}\sum_{g=1}^{G} h(\boldsymbol{\theta}^{(g)}),$$

where $g = 1, \ldots, G$ are values from the simulated Markov chain with G equal to the number of values collected after discarding the initial 'burn-in' iterations and after any thinning of the chain.

B.3 Convergence of Markov chains

MCMC methods require the use of diagnostics to assess whether the iterative simulations have reached the equilibrium distribution of the Markov chain. Sampled chains need to be run for an initial burn-in period until they can be assumed to provide approximately correct samples from the posterior distribution of interest (Lawson, 2008). Gelman and Rubin (1992) propose a general approach to monitoring convergence of MCMC simulation of c parallel chains, based on a comparison of between- and within-chain variances given that $c > 1$. Convergence of the iterative simulation is monitored by estimating the factor by which the scale of the current distribution might be reduced if the chains were continued in the limit $n \to \infty$, where n is the length of each chain. The potential scale reduction is estimated by

$$\hat{R} = \sqrt{\frac{\hat{\sigma^2}}{W}},$$

where $\hat{\sigma^2}$ is the marginal posterior variance and can be estimated by a weighted average of B and W, the between- and within- chain variances:

$$\hat{\sigma^2} = \frac{n-1}{n}W + \frac{1}{n}B.$$

$\hat{R}$ tends to 1 in the limit as $n \to \infty$ so values of $\hat{R}$ much greater than 1 suggest the chains need to be run for longer.

B.4 Deviance information criterion

Attainment of convergence of MCMC algorithms does not necessarily imply good models. For Bayesian hierarchical models, goodness-of-fit can be measured by the Akaike and Bayesian information criteria:

$$AIC = D(\hat{\boldsymbol{\theta}}) + 2p$$

$$BIC = D(\hat{\boldsymbol{\theta}}) + p \log n,$$

where $D(\hat{\boldsymbol{\theta}})$ is the deviance, p is the number of parameters and n is the number of data points. One difficulty with the AIC or BIC in models with random effects is that it is hard to decide how many parameters are included within the model (Lawson, 2008). The effective number of parameters in a hierarchical model may well be less than the total number of model parameters, due to the borrowing of strength across random effects (Banerjee et al., 2004). The difference between the posterior mean deviance $\bar{D}$ and the deviance at the posterior means $D(\bar{\boldsymbol{\theta}})$,

$$p_D = E_{\theta|y}[D] - D(E_{\theta|y}[\boldsymbol{\theta}]) = \bar{D} - D(\bar{\boldsymbol{\theta}}),$$

represents the effect of model fitting and is used as a measure of the effective number of parameters (p_D) of a Bayesian model (Gelman et al., 2004). p_D can be thought of as the number of 'unconstrained' parameters in the model where a parameter counts as 1 if it is estimated with no constraints or prior information; 0 if it is fully constrained or if all the information about the parameter comes from the prior distribution. If both the data and the prior distributions are informative, the parameter will contribute an intermediate value to p_D.

Hence, a goodness of fit measure widely used in hierarchical Bayesian modelling is the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002). The DIC is calculated as

$$DIC = \bar{D} + p_D = 2\bar{D} - D(\bar{\boldsymbol{\theta}}),$$

with smaller values of DIC indicating a better-fitting model. As with the AIC, the DIC is a measure of model fit or adequacy with a penalty for model complexity (p_D). An advantage of the DIC is that it can be easily calculated from MCMC simulation samples.

The idea is that models with smaller DIC should be preferred to models with larger DIC. The absolute size of DIC is not relevant, only differences in DIC are important.

However, according to the WinBUGS documentation (Spiegelhalter, 2008), it is difficult to say what would constitute an important difference in DIC. Very roughly, differences of more than 10 might definitely rule out the model with the higher DIC, differences between 5 and 10 are substantial, but if the difference in DIC is, say, less than 5, and the models make very different inferences, then it could be misleading just to report the model with the lowest DIC.

B.5 Posterior predictive distributions

When assessing complex Bayesian models, it can be useful to use the posterior predictive distribution as a reference distribution for comparison to the data (Gelman et al., 1996). After the data $\mathbf{y}$ have been observed, predictions of unobserved, future data $\tilde{\mathbf{y}}$ can be obtained. The distribution of $\tilde{\mathbf{y}}$ is called the posterior prediction distribution (Gelman et al., 2004), posterior because it is conditional on the observed $\mathbf{y}$ and predictive because it is a prediction for an observable $\tilde{\mathbf{y}}$:

$$p(\tilde{\mathbf{y}}|\mathbf{y}) = \int_{\boldsymbol{\theta}} p(\tilde{\mathbf{y}}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}.$$

Samples from the posterior predictive distribution of new data $\tilde{\mathbf{y}}$ are obtained using the current set of parameters, $\boldsymbol{\theta}$, and hyperparameters, $\boldsymbol{\vartheta}$, estimated from the MCMC chains. Samples from the posterior distribution of the parameters, given the observations, $p(\boldsymbol{\theta}, \boldsymbol{\vartheta}|\mathbf{y})$, are used to draw the predictive data $\tilde{\mathbf{y}}$ given $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ from the data distribution $p(\tilde{\mathbf{y}}|\boldsymbol{\vartheta}, \boldsymbol{\theta})$. The distribution of estimated values can then be compared to observed values.

Appendix C

WinBUGS code

WinBUGS code for fixed and mixed effects models M1-M5 (Chapter 5, Section 5.5).

```
#M1 - fixed effects
model
{
for (s in 1 : regions) {
  for (t in 1 : time) {
  #Negative binomial likelihood for observed counts
    cases[s,t] ~ dnegbin(p[s,t],kappa)
p[s,t]<-kappa/(kappa+mu[s,t])
    log(mu[s,t]) <- log(e[s,t])+alpha+delta[month[t]]+beta1*precip[s,t]+beta2*temp[s,t]+beta3*nino[t]
    +gamma1*alt[s]+gamma2*dens[s,t]
}
#Prior distribution for the scale parameter
kappa~dgamma(0.5,0.0005)
#Prior distribution for the intercept
alpha ~ dnorm(0.0,1.0E-6)
#Prior distributions for the month effect (annual cycle)
delta[1]<-0
for (i in 2:12)
{
delta[i] ~ dnorm(0.0,1.0E-6)
}
#Prior distributions for climate and non-climate covariates
beta1 ~ dnorm(0.0,1.0E-6)
beta2 ~ dnorm(0.0,1.0E-6)
beta3 ~ dnorm(0.0,1.0E-6)
gamma1 ~ dnorm(0.0,1.0E-6)
gamma2 ~ dnorm(0.0,1.0E-6)
#Hyperprior distributions on inverse variance parameter of random effect
tau.phi~dgamma(0.5, 0.0005)
```

```
}

#M2 - mixed effects spatially unstructured random effect
model
{
for (s in 1 : regions) {
  for (t in 1 : time) {
  #Negative binomial likelihood for observed counts
    cases[s,t] ~ dnegbin(p[s,t],kappa)
p[s,t]<-kappa/(kappa+mu[s,t])
    log(mu[s,t]) <- log(e[s,t])+alpha+delta[month[t]]+beta1*precip[s,t]+beta2*temp[s,t]+beta3*nino[t]
    +gamma1*alt[s]+gamma2*dens[s,t]+Phi[s]
}
#Prior distributions for the uncorrelated heterogeneity
Phi[s] ~ dnorm(0,tau.phi)
}
#Prior distribution for the scale parameter
kappa~dgamma(0.5,0.0005)
#Prior distribution for the intercept
alpha ~ dnorm(0.0,1.0E-6)
#Prior distributions for the month effect (annual cycle)
delta[1]<-0
for (i in 2:12)
{
delta[i] ~ dnorm(0.0,1.0E-6)
}
#Prior distributions for climate and non-climate covariates
beta1 ~ dnorm(0.0,1.0E-6)
beta2 ~ dnorm(0.0,1.0E-6)
beta3 ~ dnorm(0.0,1.0E-6)
gamma1 ~ dnorm(0.0,1.0E-6)
gamma2 ~ dnorm(0.0,1.0E-6)
#Hyperprior distributions on inverse variance parameter of random effect
tau.Phi~dgamma(0.5, 0.0005)
}

#M3 - mixed effects spatially structured random effect
model
{
for (s in 1 : regions) {
  for (t in 1 : time) {
  #Negative binomial likelihood for observed counts
    cases[s,t] ~ dnegbin(p[s,t],kappa)
p[s,t]<-kappa/(kappa+mu[s,t])
    log(mu[s,t]) <- log(e[s,t])+alpha+delta[month[t]]+beta1*precip[s,t]+beta2*temp[s,t]+beta3*nino[t]
    +gamma1*alt[s]+gamma2*dens[s,t]+Upsilon[s]
}
}
```

```
#CAR prior distribution for the spatially correlated heterogeneity
Upsilon[1:regions] ~ car.normal(adj[], weights[], num[], tau.Upsilon)
#Prior distribution for the scale parameter
kappa~dgamma(0.5,0.0005)
#Improper uniform prior distribution for the intercept
alpha ~ dflat()
#Prior distributions for the month effect (annual cycle)
delta[1]<-0
for (i in 2:12)
{
delta[i] ~ dnorm(0.0,1.0E-6)
}
#Prior distributions for climate and non-climate covariates
beta1 ~ dnorm(0.0,1.0E-6)
beta2 ~ dnorm(0.0,1.0E-6)
beta3 ~ dnorm(0.0,1.0E-6)
gamma1 ~ dnorm(0.0,1.0E-6)
gamma2 ~ dnorm(0.0,1.0E-6)
#Hyperprior distributions on inverse variance parameter of random effect
tau.Upsilon~dgamma(0.5, 0.0005)
}

#M4 - mixed effects convolution prior
model
{
for (s in 1 : regions) {
  for (t in 1 : time) {
  #Negative binomial likelihood for observed counts
    cases[s,t] ~ dnegbin(p[s,t],kappa)
p[s,t]<-kappa/(kappa+mu[s,t])
    log(mu[s,t]) <- log(e[s,t])+alpha+delta[month[t]]+beta1*precip[s,t]+beta2*temp[s,t]+beta3*nino[t]
    +gamma1*alt[s]+gamma2*dens[s,t]+phi[s]+nu[s]
}
#Prior distributions for the uncorrelated heterogeneity
phi[s] ~ dnorm(0,tau.phi)
}
#CAR prior distribution for the spatially correlated heterogeneity
nu[1:regions] ~ car.normal(adj[], weights[], num[], tau.nu)
#Prior distribution for the scale parameter
kappa~dgamma(0.5,0.0005)
#Improper uniform prior distribution for the intercept
alpha ~ dflat()
#Prior distributions for the month effect (annual cycle)
delta[1]<-0
for (i in 2:12)
{
delta[i] ~ dnorm(0.0,1.0E-6)
}
```

```
#Prior distributions for climate and non-climate covariates
beta1 ~ dnorm(0.0,1.0E-6)
beta2 ~ dnorm(0.0,1.0E-6)
beta3 ~ dnorm(0.0,1.0E-6)
gamma1 ~ dnorm(0.0,1.0E-6)
gamma2 ~ dnorm(0.0,1.0E-6)
#Hyperprior distributions on inverse variance parameter of random effects
tau.phi~dgamma(0.5, 0.0005)
tau.nu~dgamma(0.5, 0.0005)
}

#M5 - convolution prior and temporally autocorrelated calendar month effect
model
{
for (s in 1 : regions) {
  for (t in 1 : time) {
  #Negative binomial likelihood for observed counts
    cases[s,t] ~ dnegbin(p[s,t],kappa)
p[s,t]<-kappa/(kappa+mu[s,t])
    log(mu[s,t]) <- log(e[s,t])+alpha+beta1*precip[s,t]+beta2*temp[s,t]+beta3*nino[t]
    +gamma1*alt[s]+gamma2*dens[s,t]+phi[s]+nu[s]+omega[month[t]]
}
#Prior distributions for the uncorrelated heterogeneity
phi[s] ~ dnorm(0,tau.phi)
}
#CAR prior distribution for the spatially correlated heterogeneity
nu[1:regions] ~ car.normal(adj[], weights[], num[], tau.nu)
#Prior distribution for the scale parameter
kappa~dgamma(0.5,0.0005)
#Improper uniform prior distribution for the intercept
alpha ~ dflat()
#Prior distributions for the autocorrelated month effect (annual cycle)
omega[1]<-0
for (i in 2:12)
{
omega[i] ~ dnorm(omega[i-1],tau.omega)
}
#Prior distributions for climate and non-climate covariates
beta1 ~ dnorm(0.0,1.0E-6)
beta2 ~ dnorm(0.0,1.0E-6)
beta3 ~ dnorm(0.0,1.0E-6)
gamma1 ~ dnorm(0.0,1.0E-6)
gamma2 ~ dnorm(0.0,1.0E-6)
#Hyperprior distributions on inverse variance parameter of random effects
tau.phi~dgamma(0.5, 0.0005)
tau.nu~dgamma(0.5, 0.0005)
tau.omega~dgamma(0.5, 0.0005)
}
```

Appendix D

Visualisation of ternary probabilistic forecasts

A new method for visualising spatial probabilistic forecasts is described. A more detailed account of this method can be found in Jupp et al. (2010). The idea is to consider each forecast, as comprising a combination of three probabilities and to represent that as a point in a triangle of barycentric coordinates. This allows a unique colour to be assigned to each forecast from a continuum of colours defined on the triangle. Colour saturation increases with information gain relative to the climatology (reference forecast). Maps can then be produced in which the forecast at each geographical location is expressed as a colour determined by a combination of three probabilities. In contrast to current methods, forecast maps created with this colour scheme convey all of the information present in the forecast.

D.1 Three category probabilistic forecasts

A probabilistic forecast consists of a set of probabilities assigned to possible outcomes. If a 'forecasting system' (e.g. multi-model ensemble climate model or a Bayesian disease prediction model) is capable of producing probabilistic forecasts over a geographical area, these forecasts can be displayed graphically in the form of a map. The information contained within a probability forecast can be summarised in terms of categories. Here, attention will be restricted to forecasts that assign probabilities to a set of three mutu-

ally exclusive and complete outcomes (e.g. low, intermediate and high risk). Category boundaries can be defined in an appropriate way for the application and do not have to be evenly spaced. However, a useful symmetry can be introduced when the terciles of the observational dataset at each spatial location define the boundaries of the categories. Seasonal climate forecasts are commonly issued in terms of tercile categories (e.g. Barnston et al., 2003; Palmer et al., 2004; Saha et al., 2006). The labels B, N, and A will be used to denote 'below normal', 'near normal', and 'above normal' values of the forecast variable. Given these categories, the forecasting system can produce probabilistic forecasts in the form of subjective probabilities p_B, p_N, p_A that a variable, y, will be in each category at the forecast time. The probability forecast can be regarded as $\mathbf{p} = (p_B, p_N, p_N)$ with the constraints $p_B + p_N + p_A = 1$ and $0 \leq p_i \leq 1, \ \forall i$. The particular forecast $\mathbf{q} = (q_B, q_N, q_A)$ corresponds to the case where the forecaster's state of knowledge is 'no better' than the historical observed distribution of $\mathbf{y}$. For example, if the forecaster had no knowledge other than the observational record, the same forecast $\mathbf{q}$ could be issued each year. When using terciles, one third of the historical observations lie in each of the categories B, N, and A and $\mathbf{q} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. In climate science this distribution is known as *climatology*. The climatology can be viewed as the reference forecast, or benchmark distribution with which all other forecasts should be compared. As explained below, a colour will be assigned to a forecast $\mathbf{p}$ by considering the difference between the forecast $\mathbf{p}$ and the climatology $\mathbf{q}$. In addition to the forecast and the climatology, it is helpful to consider the analogous observation vector $\mathbf{o}$ at a given time and place. This represents the measured value of the variable y. For example, $\mathbf{o} = (1, 0, 0)$ when the observations lie in category B, $\mathbf{o} = (0, 1, 0)$ when the observations lie in category N, and $\mathbf{o} = (0, 0, 1)$ when the observations lie in category A. In order to visualize a probabilistic forecast $\mathbf{p}$ it will be helpful to move to barycentric coordinates and to define a measure of how similar the forecast $\mathbf{p}$ is to the reference forecast $\mathbf{q}$. The symmetric 'tercile' case $\mathbf{q} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is considered in the examples below but the more general case $\mathbf{q} \neq (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is applied to visualise probabilistic forecasts of dengue risk in Chapter 6 of this thesis.

D.2 Barycentric coordinates

Three category probabilistic forecasts have only two degrees of freedom because of the constraint that they should sum to one. Each forecast vector can be visualised as a

point in an equilateral triangle. In statistics, this visualisation technique is known as a plot in 'barycentric coordinates' (coordinates defined by the vertices of a simplex, e.g. a triangle). In applied sciences, the equivalent term 'ternary phase diagram' is used when displaying the relative proportions of three components in a mixture. Figure D.1a compares a probabilistic forecast $\mathbf{p}$ with the reference forecast $\mathbf{q}$. Note that a general ternary forecast $\mathbf{p}$ can correspond to any point in the triangle. The reference forecast $\mathbf{q}$ corresponds to the centre of the triangle (when dealing with tercile categories) and $\mathbf{o}$ corresponds to one of the corners. The general case $\mathbf{q} \neq (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is similar except that the reference forecast $\mathbf{q}$ would not lie at the centre.

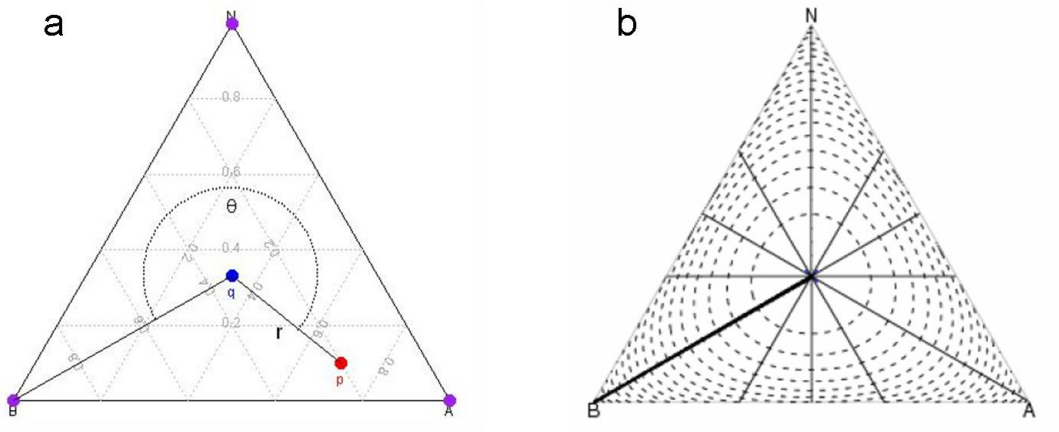


Figure D.1: (a) Triangle representing ternary forecast space in barycentric coordinates. Grey lines indicate contours of constant p_B, p_N and p_A. The reference forecast $\mathbf{q} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ lies at the centre of the plot (shown in blue). An arbitrary ternary forecast $\mathbf{p} = (0.2, 0.1, 0.7)$ is shown in red. The ternary observation $\mathbf{o}$ can take one of three possible values corresponding to the corners of the triangle (shown in purple). The angle θ is a measure of the dominant category in the ternary forecast $\mathbf{p}$. (b) Proposed coordinate system. Solid lines - contours of constant information gain $H(\mathbf{p}; \mathbf{q})$ (Eqn. 6.1). $H(\mathbf{p}; \mathbf{q})$ ranges from 0 (at the centre) to 1 (at the corners). Dashed lines - contours of constant dominant category $\theta(\mathbf{p}; \mathbf{q})/2\pi$. $\theta(\mathbf{p}; \mathbf{q})/2\pi$ is zero on the line from the centre to the bottom left and ranges from 0 to 1 moving clockwise.

D.3 Conventional practice in climate science

In seasonal climate forecasting, a discrete set of colours is often assigned to the most likely category in the ternary forecast $\mathbf{p}$. For example, Figure D.2 illustrates the 'cate-

gorical' discretisation of ternary forecast space used by the EURO-BRazilian Initiative for improving South America (EUROBRISA)[1] project. In this case, colours are assigned based on the following definitions

- 1 (Dry): ($p_B > \frac{2}{5}$ and $p_N < \frac{1}{3}$ and $p_A < \frac{1}{3}$).
- 2 (Dry or normal): ($p_B > \frac{1}{3}$ and $p_N > \frac{2}{5}$) or ($p_B > \frac{2}{5}$ and $p_N > \frac{1}{3}$).
- 3 (Normal): ($p_B < \frac{1}{3}$ and $p_N > \frac{2}{5}$ and $p_A < \frac{1}{3}$).
- 4 (Wet or normal): ($p_N > \frac{1}{3}$ and $p_A > \frac{2}{5}$) or ($p_N > \frac{2}{5}$ and $p_A > \frac{1}{3}$).
- 5 (Wet): ($p_B < \frac{1}{3}$ and $p_N < \frac{1}{3}$ and $p_A > \frac{2}{5}$).

The use of barycentric coordinates clearly highlights that there is a region of ternary forecast space (here coloured grey) that is omitted in the mathematical definition above. This region of ternary forecast space, at the base of the triangle, corresponds to ternary forecasts in which category N is assigned a low probability, but the outlying categories B and A are assigned relatively high probability. Figure D.2 illustrates a geometric example of this in barycentric coordinates. In this example, the ternary forecasts $\mathbf{p} = (1, 0, 0)$ (i.e. the forecasting system was certain that the July-August-September season would be dry) and $\mathbf{p} = (0.40, 0.30, 0.30)$ (i.e. forecast is barely different from the reference forecast $\mathbf{q} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$) would both be assigned the colour red by this categorical scheme even though they are clearly very different forecasts. Current methods for assigning colours to forecasts lose information by discretising ternary forecast space. The same colour is assigned to more than one ternary forecast. A further difficulty is that such colour assignments do not convey a consistent sense of how much *information gain* there is in the forecast.

D.4 Comparing forecasts with the climatology

In the case of tercile categories, where the reference forecast is $\mathbf{q} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, the centre of the ternary phase diagram represents a minimal state of knowledge while the corners represent absolute certainty in the forecasts. A measure of the information gain in a

[1] http://eurobrisa.cptec.inpe.br

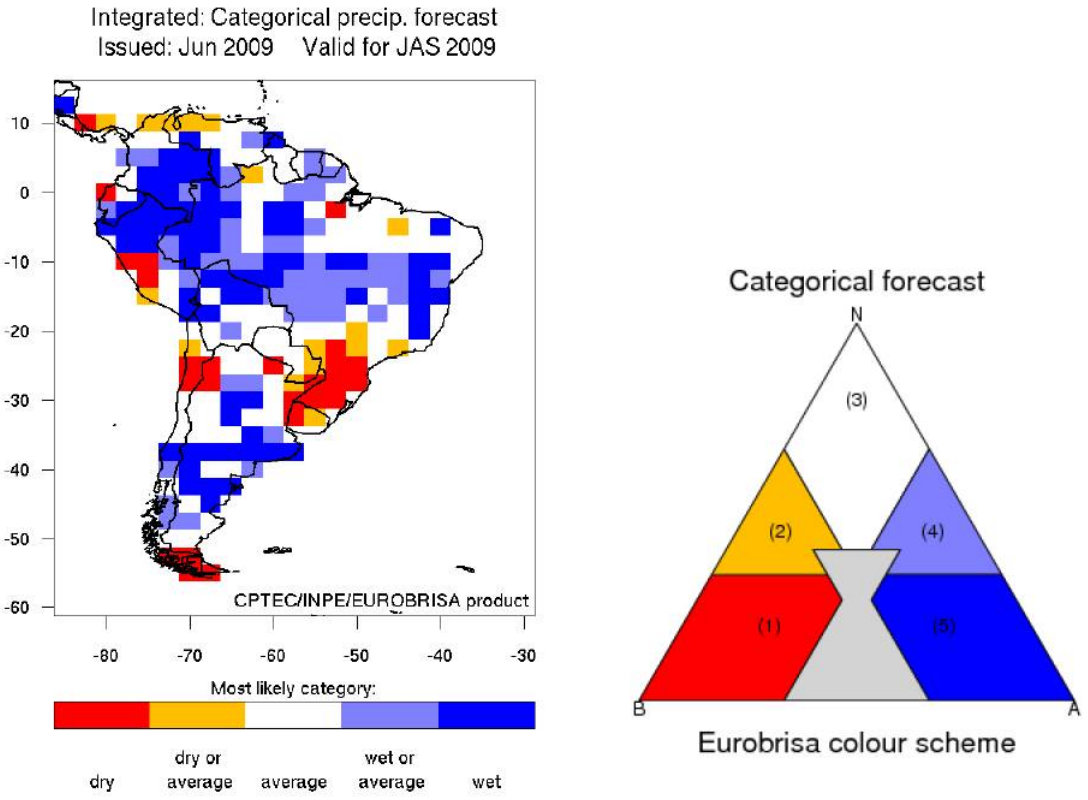


Figure D.2: A EUROBRISA ternary forecast visualised with a categorical colour scheme.

probabilistic forecast, or equivalently of the 'distance' between $\mathbf{p}$ and $\mathbf{q}$ is desired. One possibility would be to consider the Euclidean distance between the two points in Figure D.1a as a measure of certainty. However, since probabilistic forecasts are regarded as measures of belief, a measure based on entropy (Jaynes and Bretthorst, 2003) is preferred. The entropy of a distribution is defined to be

$$E(\mathbf{p}) = -\sum_i p_i \log p_i \tag{D.1}$$

where $p_i \log p_i = 0$ if $p_i = 0$. Equation D.1 is a measure of the uniformity of the distribution $\mathbf{p}$ and is maximized when all $p_i = 1/3$. It follows that a measure of the non–uniformity of $\mathbf{p}$ is the negative entropy (sometimes called the 'negentropy') $-E(\mathbf{p})$. Similarly, a standard information–theory measure of the difference between two distributions is given by the Kullback–Leibler divergence

$$K(\mathbf{p};\mathbf{q}) = \sum_i p_i \log \frac{p_i}{q_i}.$$

Consider a measure of information gain to be:

$$H(\mathbf{p};\mathbf{q}) = \frac{-1}{\log(\min(q_i^{-1}))} \sum_{i\in\{B,N,A\}} p_i \log \frac{p_i}{q_i}. \tag{D.2}$$

This can be interpreted either as a scaled version of the Kullback–Leibler divergence between $\mathbf{p}$ and $\mathbf{q}$, or as a scaled version of the (negative) entropy of $\mathbf{p}$ relative to $\mathbf{q}$. Contours of constant $H(\mathbf{p};\mathbf{q})$ are plotted in the ternary phase diagram in Figure D.1b. In the case shown, for which the categories are defined by terciles, i.e. $\mathbf{q} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, the centre of the triangle corresponds to a information gain of 0 (i.e. $H(\mathbf{q};\mathbf{q}) = 0$) and the corners (furtherest point from climatology) correspond to a information gain of 1 (i.e. $H(\mathbf{o};\mathbf{q}) = 1$). Note that $H(\mathbf{p};\mathbf{q})$ displays rotational symmetry in barycentric coordinates when the categories are defined by terciles. However, the definition of Equation D.2 also applies in the asymmetric case $\mathbf{q} \neq (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Using the idea of the most likely tercile (Fig. D.4a), the continuous angular measure $\theta(\mathbf{p};\mathbf{q})$ (Fig. D.1a) is referred to as the dominant category. The dominant category $\theta(\mathbf{p};\mathbf{q})$ measures the angle in barycentric coordinates of the forecast $\mathbf{p}$ with respect to an origin at the climatological forecast $\mathbf{q}$. Therefore, the information gain $H(\mathbf{p};\mathbf{q})$ and the dominant category $\theta(\mathbf{p};\mathbf{q})$ define a coordinate system for ternary forecast space (see Fig. D.1.b).

D.5 A new colour scheme for ternary forecasts

Colour is a natural way to represent probability forecasts in a single spatial map. It would be useful to assign a different shade of colour to each of the three categories B, N and A whilst also expressing the degree of certainty in the probabilistic forecast (encoded in the measure $H(\mathbf{p};\mathbf{q})$). In the red–green–blue (RGB) representation, a colour is represented by a set of three numbers in the range $[0, 1]$ corresponding to the brightness of the three primary colours. Thus RGB=(1,0,0) is bright red, RGB=(0,1,0) is bright green, RGB=(1,1,1) is white and RGB=(0,0,0) is black. A simple way to assign colours to forecasts would be to set RGB=(p_B, p_N, p_A). Alternatively, colours can be assigned using the hue–saturation–value (HSV) representation of colour. Geometrically, the RGB system defines a unit cube in colour space in which shades of grey occur on a 'grey line' running from the black corner RGB=(0,0,0) to the white corner RGB=(1,1,1). The HSV system describes colours as points in a cylindrical coordinate system with this grey line as its axis. The value $v \in [0, 1]$ is a measure of distance parallel to the axis (the grey

line), saturation $s \in [0, 1]$ is a measure of distance perpendicular to the axis and hue $h \in [0, 1]$ is an angular measure around the axis (see Fig D.3[2]).

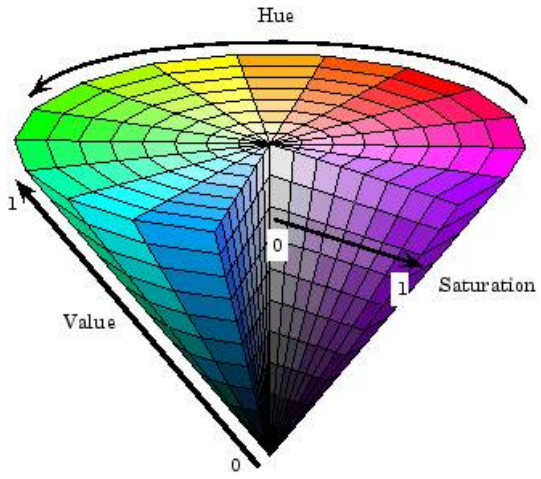


Figure D.3: HSV colour cone.

Here a colour is assigned to the forecast $\mathbf{p}$ by associating the dominant category with the hue and the information gain with the saturation:

$$\begin{aligned} \text{hue} &= \theta(\mathbf{p}; \mathbf{q})/2\pi \\ \text{saturation} &= H(\mathbf{p}; \mathbf{q}) \\ \text{value} &= 1. \end{aligned} \tag{D.3}$$

Figure D.4a shows how colours are assigned to forecasts by Equation D.3. In this case, value is set to one which causes the reference forecast to be white. The reference forecast becomes darker (grey) as value tends to zero and black when value=0. Note that a piecewise linear function has been applied to the assignment of hues that expands the yellow portion of the colour space and diminishes the green portion (see Jupp et al., 2010 for more details). Therefore, a high probability of category N is depicted by yellow (rather than green). This makes the colors more easily distinguishable to users with Daltonism; colour-blindness of the red-green type (Stephenson, 2005). Figure D.4b shows the relationship between the colour assigned to each ternary forecast and the barplot associated with each ternary forecast.

[2] http://www.mathworks.com/access/helpdesk/help/toolbox/images/hsvcone.gif [accessed 5 July 2010].

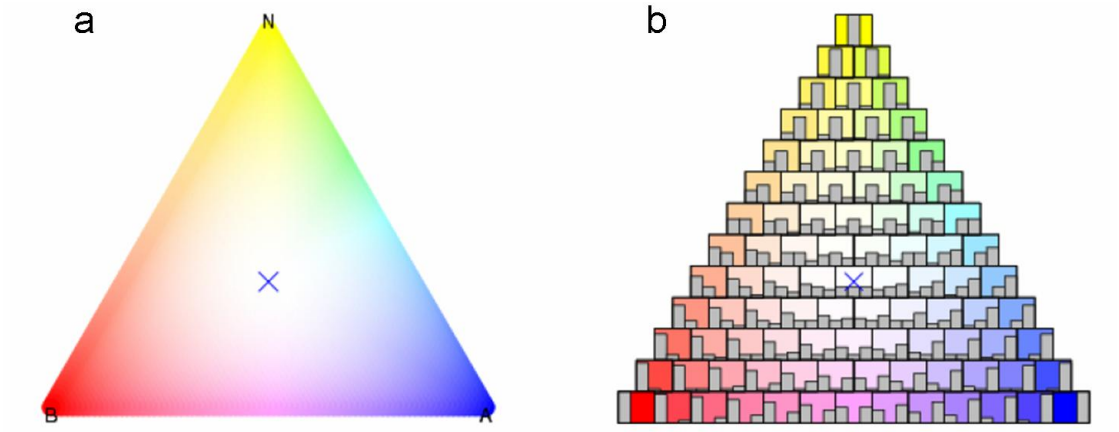


Figure D.4: (a) 'Most likely category' in barycentric coordinates. (b) Assigning colours to ternary probabilistic forecasts according to Equation D.3. (c) Relationship between the bar plot associated with each ternary forecast and the colour associated with each forecast shown in barycentric coordinates.

Glossary of Notation

A	above normal
a	hits
a_{sr}	adjacency weights
B	between-chain variance
B	below normal
b	false alarms
$b(\cdot)$	GLM function
c	misses
c	number of chains
$c(\cdot)$	GLM function
$Cov[\cdot]$	Covariance
$D(\cdot)$	(residual) deviance
d_i	contribution of the ith observation to the deviance
d	correct rejections
$E[\cdot]$	expected value
$E(\cdot)$	entropy
e_{st}	expected dengue count in microregion s and time t
g	simulated Markov chain iteration after discarding 'burn-in' and thinning
$g(\cdot)$	link function
$\mathrm{Ga}(\cdot)$	gamma probability density function
$H(\cdot)$	measure of information gain
h	hue
h_{ii}	diagonal elements of hat matrix $\mathbf{H}$ (leverages)
H_0	null hypothesis
H_1	alternative hypothesis

j	parameter index, $j = 1, \ldots, p$
$K(\cdot)$	Kullback–Leibler divergence
k	simulated Markov chain iteration
$k(s)$	geographic zone index, $k(s) = 1, \ldots, 8$
$L(\cdot)$	likelihood function
$l(\cdot)$	log-likelihood function
N	near normal
$\mathrm{N}(\cdot)$	normal (Gaussian) probability density function
n	total number of observations $n = ST$
$\mathrm{NegBin}(\cdot)$	negative binomial probability density function
o_A	binary indicator of above normal observation
o_B	binary indicator of below normal observation
o_N	binary indicator of near normal observation
p	total number of parameters
$p(\cdot)$	probability density function
p_D	effective number of parameters
p_A	probability of above normal
p_B	probability of below normal
p_N	probability of near normal
$p_{st^*(t)}$	population in microregion s and time t^*
$\mathrm{Pois}(\cdot)$	Poisson probability density function
q_A	historical frequency of above normal
q_B	historical frequency of below normal
q_N	historical frequency of near normal
$\hat{R}$	potential scale reduction factor
R^2	coefficient of determination
R^2_D	deviance explained by the model (pseudo-R^2)
$r_{(D)i}$	contribution of the ith observation to the residual deviance
$r_{(SD)i}$	Studentised deviance residuals
r	correlation
s	spatial locations index (microregions), $s = 1, \ldots, S$
s	saturation
s^2_y	sample variance of y

T	likelihood ratio test-statistic
t	time period index (month), $t = 1, \ldots, T$
$t^*(t)$	time period index (year), $t = 1, \ldots, T^*$
$t'(t)$	calendar month index, $t'(t) = 1, \ldots, 12$
$\mathrm{U}(\cdot)$	Uniform probability density function
$V(\cdot)$	variance function
v	value
$Var[\cdot]$	variance
W	within-chain variance
w_i	weights
w_{1s}	average altitude in microregion s
w_{2st}	population density in microregion s and time t
x_i	explanatory variable
x_{1st}	3-month averaged rainfall varying in space s and time t
x_{2st}	3-month averaged temperature varying in space s and time t
x_{3st}	3-month averaged ONI varying in time t
y_i	response variable
$\bar{y}$	sample mean of y
y_{st}	dengue count in microregion s and time t
$\tilde{y_{st}}$	future observed dengue count in microregion s and time t
z_i	adjusted dependent variable
α	intercept
β_j	parameter associated with climate covariates x_{jst}
β_{jk}	parameter associated with climate covariates x_{jst} in zone k
$\Gamma(\cdot)$	gamma function
γ_0	parameter associated with lagged dengue risk term $\log(\frac{y_{st-3}}{e_{st-3}})$
γ_j	parameter associated with non-climate covariates w_{jst}
$\delta_{t'(t)}$	parameter for the month factor in South East Brazil, $t'(t) = 2, \ldots, 12$
$\delta_{1t'(t)}$	parameter for the month factor in Brazil, $t'(t) = 2, \ldots, 12$
$\delta_{2k(s)}$	parameter for the zone factor in Brazil, $k(s) = 2, \ldots, 8$
$\delta_{3k(s)t'(t)}$	parameter for interaction between zone and month factor in Brazil
ζ	shape parameter

η	inverse scale parameter
η_i	linear predictor
θ	statistical model parameter
$\theta(\cdot)$	measure of dominant category
ϑ	hyperparameter
κ	scale parameter
κ^{-1}	overdispersion parameter
Λ_{st}	spatio-temporal random effects
λ_i	canonical parameter
μ_i	mean or fitted values
μ_{st}	mean dengue count in microregion s and time t
π	overall observed dengue risk for Brazil
ϖ	level of significance
ρ_{st}	relative dengue risk in microregion s and time t
σ^2	variance
σ	standard deviation
σ^2_Φ	variance for spatially unstructured random effect
σ^2_ϕ	variance for heterogeneity component of convolution prior
σ^2_Υ	variance for spatially structured random effect
σ^2_υ	variance for cluster component of convolution prior
σ^2_ω	variance for temporally correlated month effect
τ_Φ	precision for spatially unstructured random effect
τ_ϕ	precision for heterogeneity component of convolution prior
τ_Υ	precision for spatially structured random effect
τ_υ	precision for variance for cluster component of convolution prior
τ_ω	precision for temporally correlated month effect
Υ_s	spatially structured random effect
υ_s	cluster component of convolution prior
Φ_s	spatially unstructured random effect
ϕ_s	heterogeneity component of convolution prior
φ	dispersion parameter
$\chi^2_k(\cdot)$	chi-squared probability density function with k degrees of freedom
Ψ_{st}	spatio-temporal fixed effects

$\omega_{t'(t)}$	temporally correlated month effect
$\mid$	statistical symbol denoting *conditional upon*
$\sim$	statistical symbol denoting *distributed as*
$\hat{}$	statistical symbol denoting an estimated quantity or parameter
diag	mathematical notation for diagonal matrix
sgn	mathematical function that extracts the sign of a real number
$\forall$	mathematical symbol denoting *or all*

Notes:

Vectors of quantities are denoted by boldface symbols e.g. $\mathbf{y}$.

Matrices of quantities are denoted by boldface and capitalised e.g. $\mathbf{H}$.

Glossary of Acronyms

AIC	Akaike Information Criterion
AUC	Area Under ROC Curve
ARM	Autoregressive model
ARIMA	Autoregressive Integrated Moving Average
ASO	August-September-October
BIC	Bayesian Information Criterion
CAR	Conditional Autoregressive
CDC	Centers for Disease Control and Prevention
CI	Credible Interval
CIMSiM	Container-Inhabiting Mosquito Simulation Model
CPC	Climate Prediction Center
CPTEC	Center for Weather Forecasting and Climate Studies
DALY	Disability-Adjusted Life Years
DATASUS	Unified Health System Database
DEMETER	Development of a European Multi-Model Ensemble Forecast System for Seasonal to Interannual Climate Prediction
DENSiM	Dengue Simulation Model
DENV	Dengue Virus
DHF	Dengue Hemorrhagic Fever
DIC	Deviance Information Criteria
DIR	Dengue Incidence Rate
DJF	December-January-February
DSS	Dengue Shock Syndrome
ECMWF	European Centre of Medium Range Forecasts
EIR	Entomological Inoculation Rate

ENSO	El Niño Southern Oscillation
ESRL	Earth System Research Laboratory
EU FP7	European Union's Seventh Framework Programme
EUROBRISA	Euro-Brazilian Initiative for Improving South American Seasonal Forecasts
FAR	False Alarm Rate
FIOCRUZ	Oswaldo Cruz Foundation
FMA	February-March-April
GCM	General Circulation Model
GEWEX	Global Energy and Water Cycle Experiment
GLM	Generalised Linear Model
GLMM	Generalised Linear Mixed Model
GPCP	Global Precipitation Climatology Project
HIV	Human Immunodeficiency Virus
HR	Hit Rate
HSV	Hue-Saturation-Value
IBGE	Brazilian Institute for Geography and Statistics
INPE	National Institute for Space Research
IPCC	Intergovernmental Panel on Climate Change
IR	Incidence Rate
IRI	International Research Institute for Climate and Society
IRLS	Iterative Re-weighted Least Squares
ITCZ	Intertropical Convergence Zone
JAS	July-August-September
LDEO	Lamont-Doherty Earth Observatory
MCMC	Markov Chain Monte Carlo
NASA	National Aeronautics and Space Administration
NCAR	National Center for Atmospheric Research
NCEP	National Centers for Environmental Prediction
NDJ	November-December-January
NDVI	Normalized Difference Vegetation Index
NOAA	National Oceanic and Atmospheric Administration
OAR	Oceanic and Atmospheric Research

ONI	Oceanic Niño Index
OND	October-November-December
PAHO	Pan American Health Organization
PC	Proportion Correct
PNCD	The National Program for the Control of Dengue
ProMED	Program for Monitoring Emerging Diseases
PSD	Physical Sciences Division
Q-Q	Quantile-Quantile
RGB	Red-Green-Blue
ROC	Relative (or Receiver) Operating Characteristic
SIDRA	IBGE Automatic Data Collection System
SINAN	Information System for Notifiable Diseases
SMR	Standardised Morbidity Ratio
SOI	Southern Oscillation Index
SON	September-October-November
SST	Sea Surface Temperature
WHO	World Health Organisation
WinBUGS	Bayesian inference Using Gibbs Sampling for Windows

References

Abeku, T., Hay, S., Ochola, S., Langi, P., Beard, B., de Vlas, S., Cox, J., 2004. Malaria epidemic early warning and detection in African highlands. Trends in Parasitology 20 (9), 400–405.

Adjuik, M., Bagayoko, M., Binka, F., Coetzee, M., Cox, J., Craig, M., Deichman, U., Don de-Savigny, F., Fraser, C., Gouws, E. Kleinschmidt, I., Lemardeley, P., Lengeler, C., leSueur, D., Omumbo, J., Snow, B., Sharp, B., Tanser, F., Teuscher, T., Touré, Y., 1998. Towards an atlas of malaria risk in Africa. First technical report of the Mapping Malaria Risk in Africa/Atlas du Risque de la Malaria en Afrique (MARA/ARMA) collaboration. Durban, MARA/ARMA, 31pp.

Adler, R. F., Susskind, J., Huffman, G. J., Bolvin, D., Nelkin, E., Chang, A., Ferraro, R., Gruber, A., Xie, P. P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Arkin, P., 2003. The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present). Journal of Hydrometeorology 4 (6), 1147–1167.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6), 716–723.

Anderson, D., Balmaseda, M., Stockdale, T., Ferranti, L., Vitart, F., Mogensen, K., Molteni, F., Doblas-Reyes, F., Vidard, A., 2007. Development of the ECMWF seasonal forecast System 3. The European Centre for Medium-Range Weather Forecasts (ECMWF) Technical Memorandum 503, Reading, UK, 56pp.

Arcari, P., Tapper, N., Pfueller, S., 2007. Regional variability in relationships between climate and dengue/DHF in Indonesia. Singapore Journal of Tropical Geography 28 (3), 251–272.

Bailey, T., 2001. Spatial statistical methods in health. Cadernos de Saúde Pública 17, 1083–1098.

Bailey, T., Carvalho, M., Lapa, T., Souza, W., Brewer, M., 2005. Modeling of under-detection of cases in disease surveillance. Annals of Epidemiology 15 (5), 335–343.

Banerjee, S., Carlin, B., Gelfand, A., 2004. Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall/CRC, 452pp.

Barnston, A., Chelliah, M., Goldenberg, S., 1997. Documentation of a highly ENSO-related SST region in the equatorial Pacific. Atmosphere Ocean 35, 367–383.

Barnston, A. G., Mason, S. J., Goddard, L., Dewitt, D. G., Zebiak, S., 2003. Multimodel ensembling in seasonal climate forecasting at IRI. Bulletin of the American Meteorological Society 84 (12), 1783–1796.

Ben, M., Yohai, V., 2004. Quantile-quantile plot for deviance residuals in the generalized linear model. Journal of Computational and Graphical Statistics 13 (1), 36–47.

Bernardinelli, L., Clayton, D., Montomoli, C., 2007. Bayesian estimates of disease maps: how important are priors? Statistics in Medicine 14 (21-22), 2411–2431.

Besag, J., 1993. Towards Bayesian image analysis. Journal of Applied Statistics 20 (5), 107–119.

Besag, J., Green, P., Higdon, D., Mengersen, K., 1995. Bayesian computation and stochastic systems. Statistical Science 10 (1), 3–41.

Besag, J., Mollié, A., 1989. Bayesian mapping of mortality rates. Bulletin of the International Statistical Institute 53 (1), 127–128.

Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics 43 (1), 1–20.

Best, N., Arnold, R., Thomas, A., Waller, L., Conlon, E., 1999. Bayesian models for spatially correlated disease and exposure data. Bayesian Statistics 6, 131–156.

Bouma, K., Dye, C., 1997. Cycles of malaria associated with El Niño in Venezuela. Journal of the American Medical Association 278 (21), 1772–1774.

Bouma, M., Poveda, G., Rojas, W., Chavasse, D., Quinones, M., Cox, J., Patz, J., 1997. Predicting high-risk years for malaria in Colombia using parameters of El Niño Southern Oscillation. Tropical Medicine & International Health 2 (12), 1122–1127.

Braga, C., Luna, C. F., Martelli, C. M., Souza, W. V., Cordeiro, M. T., Alexander, N., Albuquerque, M. D., Júnior, J. C., Marques, E. T., 2009. Seroprevalence and risk factors for dengue infection in socio-economically distinct areas of Recife, Brazil. Acta Tropica 113 (3), 234–240.

Braga, I. A., Valle, D., 2007. Aedes aegypti: histórico do controle no Brasil (Aedes aegypti: history of control in Brazil). Epidemiologia e Serviços de Saúde 16 (2), 113–118.

Breslow, N., Clayton, D., 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88 (421), 9–25.

Bridgman, H. A., Oliver, J. E., 2006. The Global Climate System: Patterns, Processes, and Teleconnections. Cambridge University Press, New York, USA, 350pp.

Brooker, S., Hay, S. I., Bundy, D. A. P., 2002. Tools from ecology: useful for evaluating infection risk models? Trends in Parasitology 18 (2), 70–74.

Brooks, S., 1998. Markov chain Monte Carlo method and its application. Journal of the Royal Statistical Society (Series D): The Statistician 47 (1), 69–100.

Brunkard, J. M., Cifuentes, E., Rothenberg, S. J., 2008. Assessing the roles of temperature, precipitation, and ENSO in dengue re-emergence on the Texas-Mexico border region. Salud Pública de México 50 (3), 227–234.

Burke, D., Carmichael, A., Focks, D., Gimes, D., Harte, J., Lele, S., Martens, P., Mayer, J., Mearns, L., Pulwarty, R., et al., 2001. Under the Weather: Climate, Ecosystems, and Infectious Disease. Committee on Climate, Ecosystems, Infectious Diseases, and Human Health, Board on Atmospheric Sciences and Climate, National Research Council, Division on Earth and Life Studies, National Research Council. Committee on Climate, Ecosystems, Infectious Diseases, and Human Health, Board on Atmospheric Sciences and Climate, National Research Council, Division on Earth and Life Studies, National Research Council, National Academy of Sciences. Washington DC, USA: The National Academy Press, 146pp.

Câmara, F., Theophilo, R., Santos, G., Pereira, S., Câmara, D., Matos, R., 2007. Regional and dynamics characteristics of dengue in Brazil: a retrospective study. Revista da Sociedade Brasileira de Medicina Tropical 40 (2), 192–196.

Câmara, F. P., Gomes, A. F., Santos, G. T., Câmara, D. C., 2009. Climate and dengue epidemics in State of Rio de Janeiro. Revista de Sociedade Brasileira de Medicina Tropical 42 (2), 137–140.

Cameron, A., Windmeijer, F., 1996. R-squared measures for count data regression models with applications to health-care utilization. Journal of Business & Economic Statistics 14 (2), 209–220.

Cameron, A. C., Trivedi, P. K., 1998. Regression Analysis of Count Data. Cambridge University Press, New York, USA, 434pp.

Cazelles, B., Chavez, M., McMichael, A. J., Hales, S., 2005. Nonstationary influence of El Niño on the synchronous dengue epidemics in Thailand. PLoS Medicine 2 (4), 313–318.

Ceccato, P., Ghebremeskel, T., Jaiteh, M., Graves, P., Levy, M., Ghebreselassie, S., Ogbamariam, A., Barnston, A., Bell, M., del Corral, J., Connor, S. J., Fesseha, I., Brantly, E. P., Thomson, M. C., 2007. Malaria stratification, climate, and epidemic early warning in Eritrea. The American Journal of Tropical Medicine and Hygiene 77 (6 Suppl), 61–68.

Chakravarti, A., Kumaria, R., 2005. Eco-epidemiological analysis of dengue infection during an outbreak of dengue fever, India. Virology Journal 2 (1), 32–38.

Chambers, J. M., Hastie, T. J., 1992. Statistical Models in S. Chapman & Hall, New York, USA.

Chambers, T., Hahn, C., Galler, R., Rice, C., 1990. Flavivirus genome organization, expression, and replication. Annual Reviews in Microbiology 44 (1), 649–688.

Chretien, J. P., Anyamba, A., Bedno, S. A., Breiman, R. F., Sang, R., Sergon, K., Powers, A. M., Onyango, C. O., Small, J., Tucker, C. J., Linthicum, K. J., 2007. Drought-associated chikungunya emergence along coastal East Africa. The American Journal of Tropical Medicine and Hygiene 76 (3), 405–407.

Clarke, T., 2002. Dengue virus: break-bone fever. Nature 416 (6882), 672–674.

Clayton, D., Kaldor, J., 1987. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43 (3), 671–681.

Coelho, C., Uvo, C., Ambrizzi, T., 2002. Exploring the impacts of the tropical Pacific SST on the precipitation patterns over South America during ENSO periods. Theoretical and Applied Climatology 71 (3), 185–197.

Coelho, C. A. S., Pezzulli, S., Balmaseda, M., Doblas-Reyes, F. J., Stephenson, D. B., 2004. Forecast calibration and combination: A simple Bayesian approach for ENSO. Journal of Climate 17, 1504–1516.

Coelho, C. A. S., Stephenson, D. B., Balmaseda, M., Doblas-Reyes, F. J., van Oldenborgh, G. J., 2006. Toward an integrated seasonal forecasting system for South America. Journal of Climate 19 (15), 3704–3721.

Coelho, C. A. S., Stephenson, D. B., Doblas-Reyes, F. J., Balmaseda, M., 2006. The skill of empirical and combined/calibrated coupled multi-model South American seasonal predictions during ENSO. Advances in Geosciences 6, 51–55.

Coelho, C. A. S., Stephenson, D. B., Doblas-Reyes, F. J., Balmaseda, M., Guetter, A., van Oldenborgh, G. J., MAR 2006. A Bayesian approach for multi-model downscaling: Seasonal forecasting of regional rainfall and river flows in South America. Meteorological Applications 13 (1), 73–82.

Confalonieri, U., Menne, B., Akhtar, R., Ebi, K., Hauengue, M., Kovats, R., Revich, B., Woodward, A., 2007. Human Health. Climate Change 2007: Impacts, adaptation and vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK, 391–431.

Connor, S., Mantilla, G., 2008. Integration of seasonal forecasts into early warning systems for climate-sensitive diseases such as malaria and dengue. In: Seasonal Forecasts, Climatic Change and Human Health, 71–84.

Cox, J., Abeku, T., 2007. Early warning systems for malaria in Africa: from blueprint to practice. Trends in Parasitology 23 (6), 243–246.

Craig, M. H., Snow, R. W., Le Sueur, D., 1999. A climate-based distribution model of malaria transmission in sub-Saharan Africa. Trends in Parasitology 15 (3), 105–111.

Crawley, M. J., 2002. Statistical Computing: An Introduction to Data Analysis using S-Plus. John Wiley & Sons Ltd, UK, 761pp.

Cullen, J. R., Chitprarop, U., Doberstyn, E. B., Sombatwattanangkul, K., 1984. An epidemiological early warning system for malaria control in northern Thailand. Bulletin of the World Health Organization 62 (1), 107–114.

Davison, A., Gigli, A., 1989. Deviance residuals and normal scores plots. Biometrika 76 (2), 211–221.

Dean, C., Lawless, J., 1989. Tests for detecting overdispersion in Poisson regression models. Journal of the American Statistical Association 84 (406), 467–472.

Depradine, C., Lovell, E., 2004. Climatological variables and the incidence of dengue fever in Barbados. International Journal of Environmental Health Research 14 (6), 429–441.

Diggle, P., Tawn, J., Moyeed, R., 1998. Model-based geostatistics. Journal of the Royal Statistical Society: Series C (Applied Statistics) 47 (3), 299–350.

Diggle, P., Thomson, M., Christensen, O., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, J. H., Boussinesq, M., Molyneux, D. H., 2007. Spatial modelling and the prediction of Loa loa risk: decision making under uncertainty. Annals of Tropical Medicine and Parasitology 101 (6), 499–509.

Doblas-Reyes, F. J., Hagedorn, R., Palmer, T. N., 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting-II. Calibration and combination. Tellus A 57 (3), 234–252.

Doblas-Reyes, F. J., Hagedorn, R., Palmer, T. N., 2006. Developments in dynamical seasonal forecasting relevant to agricultural management. Climate Research 33 (1), 19–26.

Draper, N. R., Smith, H., 1998. Applied Regression Analysis, Third Edition. John Wiley & Sons, Inc, USA, 706pp.

Ebi, K., 2009. Malaria Early Warning Systems. In: Biometeorology for Adaptation to Climate Variability and Change, 49–74.

Eisen, R., Enscore, R., Biggerstaff, B., Reynolds, P., Ettestad, P., Brown, T., Pape, J., Tanda, D., Levy, C., Engelthaler, D., Cheek, J., Bueno, R., Targhetta, J., Montenieri1, J. A., Gage,

K. L., 2007a. Human plague in the southwestern United States, 1957-2004: Spatial models of elevated risk of human exposure to Yersinia pestis. Journal of Medical Entomology 44 (3), 530–537.

Eisen, R., Reynolds, P., Ettestad, P., Brown, T., Enscore, R., Biggerstaff, B., Cheek, J., Bueno, R., Targhetta, J., Montenieri, J., Gage, K. L., 2007b. Residence-linked human plague in New Mexico: A habitat-suitability model. The American Journal of Tropical Medicine and Hygiene 77 (1), 121–125.

Elliott, P., Wartenberg, D., 2004. Spatial epidemiology: current approaches and future challenges. Environmental Health Perspectives 112 (9), 998.

Enscore, R. E., Biggerstaff, B. J., Brown, T. L., Fulgham, R. E., Reynolds, P. J., Engelthaler, D. M., Levy, C. E., Parmenter, R. R., Montenieri, J. A., Cheek, J. E., Grinnell, R. K. Ettestad, P. J., Gage, K. L., 2002. Modeling relationships between climate and the frequency of human plague cases in the southwestern United States, 1960-1997. The American Journal of Tropical Medicine and Hygiene 66 (2), 186–196.

Epstein, P., 2001. Climate change and emerging infectious diseases. Microbes and Infection 3 (9), 747–754.

Everson, R., Fieldsend, J., 2006. Multi-class ROC analysis from a multi-objective optimisation perspective. Pattern Recognition Letters 27 (8), 918–927.

Faraway, J., 2006. Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models. Chapman & Hall/CRC, Boca Raton, Florida, USA, 331pp.

Favier, C., Degallier, N., Dubois, M. A., 2005. Dengue epidemic modelling: stakes and pitfalls. Asia Pacific Biotech News 9 (22), 1191–1194.

Favier, C., Degallier, N., Rosa-Freitas, M. G., Boulanger, J. P., Lima, J. R. C., Luitgards-Moura, J. F., Menkès, C. E., Mondet, B., Oliveira, C., Weimann, E. T. S., Tsouris, P., 2006. Early determination of the reproductive number for vector-borne diseases: the case of dengue in Brazil. Tropical Medicine & International Health 11 (3), 332–340.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters 27 (8), 861–874.

Feddersen, H., Andersen, U., 2005. A method for statistical downscaling of seasonal ensemble predictions. Tellus A 57 (3), 398–408.

Flannery, B., Pereira, M. M., Velloso, L. F., Carvalho, C. C., De Codes, L. G., Orrico, G. S., Dourado, C. M., Riley, L. W., Reis, M. G., Ko, A. I., 2001. Referral pattern of leptospirosis cases during a large urban epidemic of dengue. The American Journal of Tropical Medicine and Hygiene 65 (5), 657–663.

Focks, D., Haile, D., Daniels, E., Mount, G., 1993a. Dynamic life table model for Aedes aegypti (Diptera: Culicidae): analysis of the literature and model development. Journal of Medical Entomology 30 (6), 1003–1017.

Focks, D., Haile, D., Daniels, E., Mount, G., 1993b. Dynamic life table model for Aedes aegypti (Diptera: Culicidae): simulation results and validation. Journal of Medical Entomology 30 (6), 1018–1028.

Focks, D. A., Daniels, E., Haile, D. G., Keesling, J. E., 1995. A simulation model of the epidemiology of urban dengue fever: literature analysis, model development, preliminary validation, and samples of simulation results. The American Journal of Tropical Medicine and Hygiene 53 (5), 489–506.

Folland, C. K., Colman, A. W., Rowell, D. P., Davey, M. K., 2001. Predictability of Northeast Brazil rainfall and real-time forecast skill, 1987–98. Journal of Climate 14 (9), 1937–1958.

Fuller, D., Troyo, A., Beier, J., 2009. El Niño Southern Oscillation and vegetation dynamics as predictors of dengue fever cases in Costa Rica. Environmental Research Letters 4 (1), 014011 8pp.

Gage, K., Burkot, T., Eisen, R., Hayes, E., 2008. Climate and vectorborne diseases. American Journal of Preventive Medicine 35 (5), 436–450.

Gagnon, A., Smoyer-Tomic, K., Bush, A., 2002. The El Niño Southern Oscillation and malaria epidemics in South America. International Journal of Biometeorology 46 (2), 81–89.

Gagnon, A. S., Bush, A. B. G., Smoyer-Tomic, K. E., 2001. Dengue epidemics and the El Niño Southern Oscillation. Climate Research 19 (1), 35–43.

Garreaud, R., Vuille, M., Compagnucci, R., Marengo, J., 2009. Present-day South American climate. Palaeogeography, Palaeoclimatology, Palaeoecology 281 (3-4), 180–195.

Gelman, A., Carlin, J., Stern, H., Rubin, D., 2004. Bayesian Data Analysis, Second Edition. Chapman & Hall/CRC, Boca Raton, Florida, USA, 668pp.

Gelman, A., Meng, X., Stern, H., 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica 6, 733–759.

Gelman, A., Rubin, D. B., 1992. Inference from iterative simulation using multiple sequences. Statistical Science 7 (4), 457–472.

German, S., German, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (6), 721–741.

Gil, A. I., Louis, V. R., Rivera, I. N. G., Lipp, E., Huq, A., Lanata, C. F., Taylor, D. N., Russek-Cohen, E., Choopun, N., Sack, R. B., Colwell, R. R., 2004. Occurrence and distribution of Vibrio cholerae in the coastal environment of Peru. Environmental Microbiology 6 (7), 699–706.

Gilks, W. R., Richardson, S., Spiegelhalter, D. J., 1996. Markov Chain Monte Carlo in Practice. Chapman & Hall/CRC, Boca Raton, Florida, USA, 486pp.

Githeko, A., Lindsay, S., Confalonieri, U., Patz, J., 2000. Climate change and vector-borne diseases: a regional analysis. Bulletin of the World Health Organization 78, 1136–1147.

Githeko, A., Ndegwa, W., 2001. Predicting malaria epidemics in the Kenyan highlands using climate data: a tool for decision makers. Global Change & Human Health 2 (1), 54–63.

Goddard, L., Mason, S., 2002. Sensitivity of seasonal climate forecasts to persisted SST anomalies. Climate Dynamics 19 (7), 619–632.

Gomez-Elipe, A., Otero, A., Van Herp, M., Aguirre-Jaime, A., 2007. Forecasting malaria incidence based on monthly case reports and environmental factors in Karuzi, Burundi, 1997 – 2003. Malaria Journal 6 (129), 10pp.

Graham, R. J., Gordon, M., McLean, P. J., Ineson, S., Huddleston, M. R., Davey, M. K., Brookshaw, A., Barnes, R. T., 2005. A performance comparison of coupled and uncoupled versions of the Met Office seasonal prediction general circulation model. Tellus A 57 (3), 320–339.

Grimm, A., 2003. The El Niño impact on the summer monsoon in Brazil: regional processes versus remote influences. Journal of Climate 16, 263–280.

Grimm, A., 2004. How do La Niña events disturb the summer monsoon system in Brazil? Climate Dynamics 22 (2), 123–138.

Grimm, A., Tedeschi, R., 2009. ENSO and extreme rainfall events in South America. Journal of Climate 22, 1589–1609.

Grover-Kopec, E., Kawano, M., Klaver, R., Blumenthal, B., Ceccato, P., Connor, S., 2005. An online operational rainfall-monitoring resource for epidemic malaria early warning systems in Africa. Malaria Journal 4 (6), 5pp.

Gubler, D., 2002a. How effectively is epidemiological surveillance used for dengue programme planning and epidemic response? Dengue Bulletin 26, 96–106.

Gubler, D., Meltzer, M., 1999. Impact of dengue/dengue hemorrhagic fever on the developing world. Advances in Virus Research 53, 35–70.

Gubler, D. J., 1998. Dengue and Dengue Hemorrhagic Fever. Clinical Microbiology Reviews 11 (3), 480–496.

Gubler, D. J., 2002b. Epidemic dengue/dengue hemorrhagic fever as a public health, social and economic problem in the 21st century. Trends in Microbiology 10 (2), 100–103.

Guzman, M. G., Kouri, G., 2003. Dengue and dengue hemorrhagic fever in the Americas: lessons and challenges. Journal of Clinical Virology 27 (1), 1–13.

Hagedorn, R., Doblas-Reyes, F. J., Palmer, T. N., 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting–I. Basic concept. Tellus A 57 (3), 219–233.

Hales, S., de Wet, N., Maindonald, J., Woodward, A., 2002. Potential effect of population and climate changes on global distribution of dengue fever: an empirical model. The Lancet 360 (9336), 830–834.

Hales, S., Weinstein, P., Souares, Y., Woodward, A., 1999. El Niño and the dynamics of vector-borne disease transmission. Environmental Health Perspectives 107 (2), 99–102.

Hales, S., Weinstein, P., Woodward, A., 1996. Dengue fever epidemics in the South Pacific: driven by El Niño Southern Oscillation? Lancet 348 (9042), 1664–1665.

Halstead, S., Deen, J., 2002. The future of dengue vaccines. Lancet 360 (9341), 1243–1245.

Halstead, S. B., 1981. The Alexander D. Langmuir Lecture. The pathogenesis of dengue. Molecular epidemiology in infectious disease. American Journal of Epidemiology 114 (5), 632–648.

Hastenrath, S., 2006. Circulation and teleconnection mechanisms of Northeast Brazil droughts. Progress in Oceanography 70 (2-4), 407–415.

Hastings, W., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57 (1), 97–109.

Hay, S. I., Simba, M., Busolo, M., Noor, A. M., Guyatt, H. L., Ochola, S. A., Snow, R. W., 2002. Defining and detecting malaria epidemics in the highlands of western Kenya. Emerging Infectious Diseases 8 (6), 555–562.

Hayden, M., Uejio, C., Walker, K., Ramberg, F., Moreno, R., Rosales, C., Gameros, M., Mearns, L., Zielinski-Gutierrez, E., Janes, C., 2010. Microclimate and Human Factors in the Divergent Ecology of Aedes aegypti along the Arizona, US/Sonora, MX Border. EcoHealth 7 (1), 64–77.

Hilbe, J., 2007. Negative Binomial Regression. Cambridge University Press, New York, USA, 264pp.

Hoaglin, D., Welsch, R., 1978. The hat matrix in regression and ANOVA. American Statistician 32 (1), 17–22.

Hoshen, M., Morse, A., 2004. A weather-driven model of malaria transmission. Malaria Journal 3 (32), 14pp.

Hunter, P. R., 2003. Climate change and waterborne and vector-borne disease. Journal of Applied Microbiology 94 (s1), 37–46.

Hurtado-Diaz, M., Riojas-Rodriguez, H., Rothenberg, S. J., Gomez-Dantes, H., Cifuentes, E., 2007. Short communication: Impact of climate variability on the incidence of dengue in Mexico. Tropical Medicine & International Health 12 (11), 1327–1337.

Jaynes, E., Bretthorst, G., 2003. Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, UK, 727pp.

Johansson, M. A., Cummings, D. A. T., Glass, G. E., 2009a. Multi-year variability and dengue - El Niño Southern Oscillation, weather, and dengue incidence in Puerto Rico, Mexico, and Thailand: a longitudinal data analysis. PLoS Medicine 6 (11), e1000168, doi:10.1371/journal.pmed.1000168.

Johansson, M. A., Dominici, F., Glass, G., 2009b. Local and Global Effects of Climate on Dengue Transmission in Puerto Rico. PLoS Neglected Tropical Diseases 3 (2), e382.

Jolliffe, I. T., Stephenson, D. B., 2003. Forecast Verification: A Practitioner's Guide in Atmospheric Science. John Wiley & Sons Ltd, UK, 240pp.

Jones, A., 2007. Seasonal ensemble prediction of malaria in africa. Ph.D. thesis, University of Liverpool.

Jones, A., Wort, U., Morse, A., Hastings, I., Gagnon, A., 2007. Climate prediction of El Niño malaria epidemics in north-west Tanzania. Malaria Journal 6 (162), 15pp.

Jones, A. E., Morse, A. P., 2010. Application and validation of a seasonal ensemble prediction system using a dynamic malaria model. Journal of Climate 23 (15), 4202–4215.

Julious, S., Nicholl, J., George, S., 2001. Why do we continue to use standardized mortality ratios for small area comparisons? Journal of Public Health 23 (1), 40–46.

Jupp, T., Lowe, R., Stephenson, D. B., Coelho, C. A. S., 2010. On the interpretation, verification and calibration of ternary probabilistic forecasts. [manuscript in preparation].

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., Joseph, D., 1996. The NCAR/NCEP 40-year reanalysis project. Bulletin of the American Meteorological Society 77 (3), 437–471.

Kelly-Hope, L., Thomson, M. C., 2008. Climate and infectious diseases. In: Seasonal Forecasts, Climatic Change and Human Health, 31–70.

Knorr-Held, L., 2000. Bayesian modelling of inseparable space-time variation in disease risk. Statistics in Medicine 19 (1718), 2555–2567.

Kovats, R. S., 2000. El Niño and human health. Bulletin of the World Health Organization 78 (9), 1127–1135.

Kovats, R. S., Bouma, M. J., Hajat, S., Worrall, E., Haines, A., 2003. El Niño and health. Lancet 362 (9394), 1481–1489.

Krzanowski, W., 1998. An Introduction to Statistical Modelling. Arnold London, UK, 252pp.

Kuhn, K., Campbell-Lendrum, D., Haines, A., Cox, J., Corvalán, C., Anker, M., 2005. Using climate to predict infectious disease epidemics. World Health Organization, Geneva, 54pp.

Kuhn, K. G., Campbell-Lendrum, D., Haines, A., Cox, J., 2004. Using climate to predict infectious disease outbreaks: A review. World Health Organization, Geneva, 55pp.

Kuno, G., 1995. Review of the factors modulating dengue transmission. Epidemiologic Reviews 17 (2), 321–335.

Lafferty, K. D., 2009. The ecology of climate change and infectious diseases. Ecology 90 (4), 888–900.

Lama, J., Seas, C., León-Barúa, R., Gotuzzo, E., Sack, R., 2004. Environmental temperature, cholera, and acute diarrhoea in adults in Lima, Peru. Journal of Health, Population and Nutrition 22, 399–403.

Lawless, J. F., 1987. Negative binomial and mixed Poisson regression. Canadian Journal of Statistics 15 (3), 209–225.

Lawson, A., 2008. Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. Chapman & Hall/CRC, Boca Raton, Florida, USA, 344pp.

Lawson, A. B., Browne, W. J., Rodeiro, C. L. V., 2003. Disease Mapping with WinBUGS and MLwiN. John Wiley & Sons Ltd, UK, 277pp.

Liebmann, B., Marengo, J., 2001. Interannual variability of the rainy season and rainfall in the Brazilian Amazon basin. Journal of Climate 14, 4308–4318.

Lowe, R., Bailey, T. C., Stephenson, D. B., Graham, R., Coelho, C. A. S., Carvalho, M. S., Barcellos, C., 2009. Climate-based dengue predictions for Brazil. In: Proceedings of StatGIS09: GeoInformatics for Environmental Surveillance, Milos, Greece, 17-19 June 2009.

Lowe, R., Bailey, T. C., Stephenson, D. B., Graham, R. J., Coelho, C. A. S., Carvalho, M. S., Barcellos, C., 2010. Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil. Computers & Geosciences, doi:10.1016/j.cageo.2010.01.008.

Lunn, D., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing 10 (4), 325–337.

Luz, P., Grinsztejn, B., Galvani, A., 2009. Disability adjusted life years lost to dengue in Brazil. Tropical Medicine & International Health 14 (2), 237–246.

Luz, P. M., Mendes, B. V. M., Codeco, C. T., Struchiner, C. J., Galvani, A. P., 2008. Time series analysis of dengue incidence in Rio de Janeiro, Brazil. The American Journal of Tropical Medicine and Hygiene 79 (6), 933–939.

Lyon, B., Barnston, A. G., 2005. ENSO and the spatial extent of interannual precipitation extremes in tropical land areas. Journal of Climate 18 (23), 5095–5109.

Mabaso, M., Vounatsou, P., Midzi, S., Da Silva, J., Smith, T., 2006. Spatio-temporal analysis of the role of climate in inter-annual variation of malaria incidence in Zimbabwe. International Journal of Health Geographics 5 (20), 9pp.

MacNab, Y., 2003. Hierarchical Bayesian modeling of spatially correlated health service outcome and utilization rates. Biometrics 59 (2), 305–316.

Mantilla, G., Oliveros, H., Barnston, A. G., 2009. The role of ENSO in understanding changes in Colombia's annual malaria burden by region, 1960 – 2006. Malaria Journal 8 (6), 11pp.

Marengo, J., Hastenrath, S., 1993. Case studies of extreme climatic events in the Amazon Basin. Journal of Climate 6, 617–627.

Martens, P., Kovats, R., Nijhof, S., De Vries, P., Livermore, M., Bradley, D., Cox, J., McMichael, A., 1999. Climate change and future populations at risk of malaria. Global Environmental Change 9, S89–S107.

Mason, I., 1979. On reducing probability forecasts to yes/no forecasts. Monthly Weather Review 107 (2), 207–211.

Mason, I., 2003. Binary Events. In: Forecast Verification: A Practitioners Guide in Atmospheric Science, 37–76.

Mason, S., Goddard, L., 2001. Probabilistic precipitation anomalies associated with ENSO. Bulletin of the American Meteorological Society 82 (4), 619–638.

Mason, S. J., Graham, N. E., 2002. Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation. Quarterly Journal of the Royal Meteorological Society 128 (584), 2145–2166.

McBride, W., Bielefeldt-Ohmann, H., 2000. Dengue viral infections; pathogenesis and epidemiology. Microbes and Infection 2 (9), 1041–1050.

McCullagh, P., Nelder, J., 1989. Generalized Linear Models, Second Edition. Chapman Hall, London, UK, 536pp.

McCulloch, C., Searle, S., 2001. Generalized, Linear, and Mixed Models. John Wiley & Sons, Inc., Canada, 358pp.

McMichael, A. J., Campbell-Lendrum, D. H., Corvalán, C. F., Ebi, K. L., Githeko, A. K., Scheraga, J. D., Woodward, A., 2003. Climate change and human health: risks and responses. World Health Organization Geneva, 322pp.

McMichael, A. J., Haines, A., Slooff, R. Kovats, S., 1996. Climate change and human health: an assessment prepared by a Task Group on behalf of the World Health Organization, the World Meteorological Organization and the United Nations Environment Programme. World Health Organization, Geneva, 297pp.

McPhaden, M., Zebiak, S., Glantz, M., 2006. ENSO as an integrating concept in earth science. Science 314 (5806), 1740–1745.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E., 1953. Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21 (6), 1087–1092.

Misra, V., Dirmeyer, P., Kirtman, B., 2003. Dynamic downscaling of seasonal simulations over South America. Journal of Climate 16, 103–117.

Mollie, A., 1996. Bayesian mapping of disease. In: Markov Chain Monte Carlo in Practice, 359–379.

Monath, T. P., 1994. Dengue: the risk to developed and developing countries. Proceedings of the National Academy of Sciences 91 (7), 2395–2400.

Montecinos, A., Díaz, A., Aceituno, P., 2000. Seasonal diagnostic and predictability of rainfall in subtropical South America based on tropical Pacific SST. Journal of Climate 13 (4), 746–758.

Morse, A. P., Doblas-Reyes, F. J., Hoshen, M. B., Hagendorn, R., Palmer, T. I. M. N., 2005. A forecast quality assessment of an end-to-end probabilistic multi-model seasonal forecast system using a malaria model. Tellus 57A (3), 464–475.

Murray, C., 1994. Quantifying the burden of disease: the technical basis for disability-adjusted life years. Bulletin of the World Health Organization 72 (3), 429–445.

Murray, C., Lopez, A., 1994. Quantifying disability: data, methods and results. Bulletin of the World Health Organization 72 (3), 481–494.

Murray, C., Lopez, A., 2002. World Health Report: 2002: Reducing risks, promoting healthy life. World Health Organization, Geneva, 248pp.

Myers, M. F., Rogers, D. J., Cox, J., Flahault, A., Hay, S. I., 2000. Forecasting disease risk for increased epidemic preparedness in public health. Advances in Parasitology 47, 309–330.

Nakhapakorn, K., Tripathi, N. K., 2005. An information value based analysis of physical and climatic factors affecting dengue fever and dengue haemorrhagic fever incidence. International Journal of Health Geographics 4 (13), 13pp.

Nelder, J. A., Wedderburn, R. W. M., 1972. Generalized linear models. Journal of the Royal Statistical Society. Series A (General) 135 (3), 370–384.

Nogueira, R. M. R., Araújo, J. M. G., Schatzmayr, H. G., 2007a. Dengue viruses in Brazil, 1986-2006. Revista Panamericana de Salud Pública 22 (5), 358–363.

Nogueira, R. M. R., da Araújo, J. M. G., Schatzmayr, H. G., 2007b. Aspects of dengue virus infections in Brazil 1986-2007. Virus Reviews and Research 12, 1–17.

Nogueira, R. M. R., Miagostovich, M. P., Lampe, E., Schatzmayr, H. G., 1990. Isolation of dengue virus type 2 in Rio de Janeiro. Memórias do Instituto Oswaldo Cruz 85 (2), 253.

Nogueira, R. M. R., Schatzmayr, H. G., de Filippis, A. M., dos Santos, F. B., da Cunha, R. V., Coelho, J. O., de Souza, L. J., Guimarães, F. R., de Araújo, E. S., De Simone, T. S., Baran, M. Teixeira, J. G., 2005. Dengue virus type 3, Brazil, 2002. Emerging Infectious Diseases 11 (9), 1376–1381.

Omumbo, J., Hay, S., Snow, R., Tatem, A., Rogers, D., 2005. Modelling malaria risk in East Africa at high-spatial resolution. Tropical Medicine & International Health 10 (6), 557–566.

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Délécluse, P., Déqué, M., Díez, E., Doblas-Reyes, F. J., Feddersen, H., Graham, R., Gualdi, S., Guérémy, J. F., Hagedorn, R., Hoshen, M., Keenlyside, N., Latif, M., Lazar, A., Maisonnave, E., Marletto, V., Morse, A. P., Orfila, B., Rogel, P., Terres, J. M., Thomson, M. C., 2004. Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). Bulletin of the American Meteorological Society 85 (6), 853–872.

Parmenter, R., Yadav, E., Parmenter, C., Ettestad, P., Gage, K., 1999. Incidence of plague associated with increased winter-spring precipitation in New Mexico. The American Journal of Tropical Medicine and Hygiene 61 (5), 814–821.

Pascual, M., Rodó, X., Ellner, S., Colwell, R., Bouma, M., 2000. Cholera dynamics and El Nino-southern oscillation. Science 289 (5485), 1766–1769.

Patz, J. A., Martens, W. J., Focks, D. A., Jetten, T. H., 1998. Dengue fever epidemic potential as projected by general circulation models of global climate change. Environmental Health Perspectives 106 (3), 147–153.

Pepe, M. S., 2004. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, USA, 302pp.

Philander, S., 1990. El Niño, La Niña, and the Southern Oscillation. Academic Press, USA, 293pp.

Pontes, R. J., Freeman, J., Oliveira-Lima, J. W., Hodgson, J., Spielman, A., 2000. Vector densities that potentiate dengue outbreaks in a Brazilian City. The American Journal of Tropical Medicine and Hygiene 62 (3), 378–383.

Poveda, G., Rojas, W., Quiñones, M., Vélez, I., Mantilla, R., Ruiz, D., Zuluaga, J., Rua, G., 2001. Coupling between annual and ENSO timescales in the malaria-climate association in Colombia. Environmental Health Perspectives 109 (5), 489–493.

Promprou, S., Jaroensutasinee, M., Jaroensutasinee, K., 2005. Climatic factors affecting dengue haemorrhagic fever incidence in Southern Thailand. Dengue Bulletin 29, 41–48.

R Development Core Team, 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL `http://www.R-project.org`

Raftery, A. E., Lewis, S. M., 1996. Implementing MCMC. In: Markov Chain Monte Carlo in Practice, 115–130.

Reiter, P., 2001. Climate change and mosquito-borne disease. Environmental Health Perspectives 109 (Suppl 1), 141–161.

Reiter, P., Lathrop, S., Bunning, M., Biggerstaff, B., Singer, D., Tiwari, T., Baber, L., Amador, M., Thirion, J., Hayes, J., Seca, C., Mendez, J., Ramirez, B., Robinson, J., Rawlings, J., Vorndam, V., Waterman, S., Gubler, D., Clark, G., Hayes, E., 2003. Texas lifestyle limits transmission of dengue virus. Emerging Infectious Diseases 9 (1), 86–89.

Rigau-Pérez, J., Clark, G., Gubler, D., Reiter, P., Sanders, E., Vance Vorndam, A., 1998. Dengue and dengue haemorrhagic fever. The Lancet 352 (9132), 971–977.

Rodó, X., Pascual, M., Fuchs, G., Faruque, A., 2002. ENSO and cholera: A nonstationary link related to climate change? Proceedings of the National Academy of Sciences 99 (20), 12901–12906.

Ropelewski, C. F., Halpert, M. S., 1987. Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. Monthly Weather Review 115 (8), 1606–1626.

Rueda, L., Patel, K., Axtell, R., Stinner, R., 1990. Temperature-dependent development and survival rates of Culex quinquefasciatus and Aedes aegypti (Diptera: Culicidae). Journal of Medical Entomology 27 (5), 892–898.

Saha, S., Nadiga, S., Thiaw, C., Wang, J., Wang, W., Zhang, Q., Van den Dool, H. M., Pan, H. L., Moorthi, S., Behringer, D., Stokes, D., Peña, M., Lord, S., White, G., Ebisuzaki, W., Peng, P., Xie, P., 2006. The NCEP climate forecast system. Journal of Climate 19, 3483–3517.

Sanchez, L., Vanlerberghe, V., Alfonso, L., Marquetti, M., Guzman, M., Bisset, J., van der Stuyft, P., 2006. Aedes aegypti larval indices and risk for dengue epidemics. Emerging Infectious Diseases 12 (5), 800–806.

Schatzmayr, H. G., Nogueira, R. M. R., Rosa, A. P. A. T., 1986. An outbreak of dengue virus at Rio de Janeiro-1986. Memórias do Instituto Oswaldo Cruz 81, 245–246.

Schreiber, K. V., 2001. An investigation of relationships between climate and dengue using a water budgeting technique. International Journal of Biometeorology 45, 81–89.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6 (2), 461–464.

Scott, T., Amerasinghe, P., Morrison, A., Lorenz, L., Clark, G., Strickman, D., Kittayapong, P., Edman, J., 2000. Longitudinal studies of Aedes aegypti (Diptera: Culicidae) in Thailand and Puerto Rico: blood feeding frequency. Journal of Medical Entomology 37 (1), 89–101.

Siqueira, J., Martelli, C., Coelho, G., Simplício, A., Hatch, D., 2005. Dengue and dengue hemorrhagic fever, Brazil, 1981-2002. Emerging Infectious Diseases 11 (1), 48–53.

Snall, T., O'Hara, R., Ray, C., Collinge, S., 2008. Climate-driven spatial dynamics of plague among prairie dog colonies. The American Naturalist 171 (2), 238–248.

Spiegelhalter, D., 2008. The BUGS project - DIC. Technical Report, MRC Biostatistics Unit, Cambridge, UK.
URL `http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/dicpage.shtml`

Spiegelhalter, D., Best, N., Carlin, B., van der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 64 (4), 583–639.

Stapp, P., Antolin, M., Ball, M., 2004. Patterns of extinction in prairie dog metapopulations: plague outbreaks follow El Niño events. Frontiers in Ecology and the Environment 2 (5), 235–240.

Stephenson, D., 2005. Comment on Color schemes for improved data graphics, by A. Light and PJ Bartlein. Eos Transactions American Geophysical Union 86 (20), 196.

Stephenson, D. B., Coelho, C. A. S., Doblas-Reyes, F. J., Balmaseda, M., MAY 2005. Forecast assimilation: a unified framework for the combination of multi-model weather and climate predictions. Tellus 57A (3), 253–264.

Strickman, D., Kittayapong, P., 2002. Dengue and its vectors in Thailand: introduction to the study and seasonal distribution of Aedes larvae. The American Journal of Tropical Medicine and Hygiene 67 (3), 247–259.

Sturtz, S., Ligges, U., Gelman, A., 2005. R2WinBUGS: a package for running WinBUGS from R. Journal of Statistical Software 12 (3), 1–16.

Tanser, F., Sharp, B., le Sueur, D., 2003. Potential effect of climate change on malaria transmission in Africa. The Lancet 362 (9398), 1792–1798.

Teixeira, M., Costa, M., Barreto, F., Barreto, M., 2009. Dengue: twenty-five years since reemergence in Brazil. Cadernos de Saúde Pública 25, 7–18.

Teixeira, M. G., Costa, M. C. N., Barreto, M. L., Mota, E., 2005. Dengue and dengue hemorrhagic fever epidemics in Brazil: what research is needed based on trends, surveillance, and control experiences? Cadernos de Saúde Pública 21, 1307–1315.

Teklehaimanot, H. D., Schwatrz, J., Teklehaimanot, A., Lipsitch, M., 2004. Alert threshold algorithms and malaria epidemic detection. Emerging Infectious Diseases 10 (7), 1220–1226.

Thammapalo, S., Chongsuwiwatwong, V., McNeil, D., Geater, A., 2005. The climatic factors influencing the occurrence of dengue hemorrhagic fever in Thailand. The Southeast Asian Journal of Tropical Medicine and Public Health 36 (1), 191–196.

Thomson, M., Obsomer, V., Kamgno, J., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Remme, J., Molyneux, D., Boussinesq, M., 2004. Mapping the distribution of Loa loa in Cameroon in support of the African Programme for Onchocerciasis Control. Filaria Journal 3 (7), 13pp.

Thomson, M. C., Connor, S. J., SEP 2001. The development of malaria early warning systems for Africa. Trends in Parasitology 17 (9), 438–445.

Thomson, M. C., Doblas-Reyes, F. J., Mason, S. J., Hagedorn, R., Connor, S. J., Phindela, T., Morse, A. P., Palmer, T. N., 2006. Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. Nature 439 (7076), 576–579.

Thomson, M. C., Mason, S. J., Phindela, T., Connor, S. J., 2005. Use of rainfall and sea surface temperature monitoring for malaria early warning in Botswana. The American Journal of Tropical Medicine and Hygiene 73 (1), 214–221.

Tipayamongkholgul, M., Fang, C., Klinchan, S., Liu, C., King, C., 2009. Effects of the El Niño-Southern Oscillation on dengue epidemics in Thailand, 1996-2005. BMC Public Health 9 (422), 15pp.

Torrence, C., Webster, P. J., 1998. The annual cycle of persistence in the El Niño/Southern Oscillation. Quarterly Journal of the Royal Meteorological Society 124 (550), 1985–2004.

Venables, W. N., Ripley, B. D., 2002. Modern Applied Statistics With S. Springer, New York, USA, 495pp.

Wakefield, J. C., Best, N. G., Waller, L., 2000. Bayesian approaches to disease mapping. In: Spatial Epidemiology: Methods and Applications, 104–127.

Watts, D., Burke, D., Harrison, B., Whitmire, R., Nisalak, A., 1987. Effect of temperature on the vector efficiency of Aedes aegypti for dengue 2 virus. The American Journal of Tropical Medicine and Hygiene 36 (1), 143–152.

Webster, D., Farrar, J., Rowland-Jones, S., 2009. Progress towards a dengue vaccine. The Lancet Infectious Diseases 9 (11), 678–687.

Wedderburn, R., 1974. Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. Biometrika 61 (3), 439–447.

Winters, A. M., Staples, J. E., Ogen-Odoi, A., Mead, P. S., Griffith, K., Owor, N., Babi, N., Enscore, R. E., Eisen, L., Gage, K. L., J., E. R., 2009. Spatial risk models for human plague in the West Nile region of Uganda. The American Journal of Tropical Medicine and Hygiene 80 (6), 1014–1022.

Worrall, E., Connor, S. J., Thomson, M. C., 2007. A model to simulate the impact of timing, coverage and transmission intensity on the effectiveness of indoor residual spraying (IRS) for malaria control. Tropical Medicine & International Health 12 (1), 75–88.

Wu, P., Guo, H., Lung, S., Lin, C., Su, H., 2007. Weather as an effective predictor for occurrence of dengue fever in Taiwan. Acta Tropica 103 (1), 50–57.

Wu, P., Lay, J., Guo, H., Lin, C., Lung, S., Su, H., 2009. Higher temperature and urbanization affect the spatial patterns of dengue fever transmission in subtropical Taiwan. Science of the Total Environment 407 (7), 2224–2233.

Zeileis, A., Kleiber, C., Jackman, S., 2008. Regression Models for Count Data in R. Journal of Statistical Software 27 (8), 1–25.

Zhou, G., Minakawa, N., Githeko, A., Yan, G., 2004. Association between climate variability and malaria epidemics in the East African highlands. Proceedings of the National Academy of Sciences of the United States of America 101 (8), 2375–2380.

Zuur, A., Ieno, E., Walker, N., Saveliev, A., Smith, G., 2009. Mixed effects models and extensions in ecology with R. Springer, New York, USA, 574pp.